COST INFORMATION-VALUE TRADE-OFF IN COVARIATE SELECTION

by

PEDER ØSTBYE

MASTER THESIS

for the degree

Master of science

(Master i modellering og dataanalyse)

Faculty of Mathematics and Natural Sciences
University of Oslo

November 2014

Det matematisk- naturvitenskapelige fakultet
Universitetet i Oslo

Preface

I am confident that mathematical statistics will enter the history as one of the major scientific discoveries of the 20th century, if not the single most important discovery. Statistics is an inevitable input to all sciences, and the present body of human knowledge relies to a large extent on statistical inferences.

What an exciting time to study statistics! I have learned from masters in the field, such as my thesis supervisor, Professor Nils Lid Hjort. I have observed major developments in the field as they have happened, such as how modern model selection methods have entered mainstream text books. I have experienced how home laptops can perform statistical analyses that just a few years ago were reserved university computers. I have also been able to see the dangers lurking in the shadow of statistical inference when not paying attention to the assumptions, and even the abuse of statistics in the pursuit of biased interests. The latter reflects the importance of statistical investigations on high stakes political debates, such as climate change.

I am very happy to have made an ε-contribution to the field of statistics by this thesis. I am thankful to whatever gifted me with sufficient mathematical talent to make statistics a part of my inquiry into various sciences. I am also thankful for the liberal and easy-to-access Norwegian education system, allowing studying to be a life long process and the possibility to pursue interdisciplinary knowledge.

Peder Østbye
Oslo, November 2014

Contents

1 Introduction and main findings **1**

2 Statistical context and existing literature **2**

3 Analytical framework and principles **5**
3.1 Regression models . . . 5
3.1.1 The structure of regression models . . . 5
3.1.2 Estimation and prediction . . . 7
3.2 Maximum likelihood estimators . . . 8
3.2.1 Maximum likelihood estimators and their properties . . . 8
3.2.2 The delta method . . . 12
3.3 Covariates and information value . . . 13
3.3.1 Reduced Kullback-Leibler distance as information value . . . 13
3.3.2 Reduced mean squared error as information value . . . 16
3.3.3 Reduction in expected loss as a general framework for information value . . . 21
3.3.4 Risk function and Bayesian analysis . . . 25
3.4 Cost of gathering covariates . . . 26
3.5 Optimization . . . 27

4 Covariate cost functions **28**
4.1 Simple linear covariate costs . . . 28
4.2 Economies of scope and sequence . . . 29
4.3 Some words on costs subject to random influences . . . 30

5 Loss functions and estimating the information-value of covariates **31**
5.1 Constructing loss functions for estimating economic information-value . . . 31
5.1.1 Economic information value versus statistical information-value . . . 31
5.1.2 The general properties of reasonable loss functions . . . 32
5.1.3 Some examples of particular loss functions and economic loss calibration . . . 33
5.1.4 Finding the loss-function parameters . . . 37
5.2 Covariates, parameter complexity and information value . . . 38
5.3 Direct estimation of expected loss in prediction settings . . . 38
5.3.1 General approach . . . 38
5.3.2 Squared prediction error as loss function . . . 39
5.3.3 Zero-one loss in two class prediction . . . 43
5.3.4 AIC in an another loss estimation perspective . . . 45
5.3.5 The usefulness of direct expected loss estimation . . . 46

5.4 A cross-validation approach to expected loss estimation in prediction settings 47
5.4.1 Estimating expected loss by cross-validation 47
5.4.2 What do we estimate when using cross validation? 48
5.4.3 Computational issues when using cross validation in estimation 50
5.4.4 The usefulness of cross-validation in expected loss estimation 51
5.4.5 Jackknife estimation 51
5.4.6 Bootstrapping 52
5.5 A FIC-inspired approach to expected loss estimation 52
5.5.1 From FIC to AFIC 53
5.5.2 Applying FIC to more general loss functions 56
5.5.3 Direct use of FIC in the case of LINEX loss 60
5.5.4 The use of FIC in prediction settings 62
5.6 Chapter summary 63

6 Loss estimation and cost information-value trade-off illustrated by a simulation experiment 64
6.1 The experiment setup 64
6.1.1 The data generating process 64
6.1.2 Foci to be considered 67
6.1.3 Loss functions considered 70
6.1.4 The information-value cost trade-off in the experiment 70
6.2 Experiment results and analysis for μ^1 71
6.2.1 The result of the experiment 71
6.2.2 Comparing the estimation methods 72
6.2.3 The information value of x_3 73
6.2.4 The value of more precise information 74
6.2.5 The value of increased amount of data 74
6.3 Experiment results for μ^2 74
6.3.1 The result of the experiment 74
6.3.2 Comparing the estimation methods 75
6.3.3 The information value of x_3 76
6.3.4 The value of more precise information 77
6.3.5 The value of increased amount of data 77
6.4 Experiment results for μ^3 77
6.4.1 The result of the experiment 77
6.4.2 Comparing the estimation methods 77
6.4.3 The information value of x_3 84
6.4.4 The value of more precise information 85
6.4.5 The value of increased amount of data 85

6.5 Chapter summary . . . 85

7 Concluding remarks and more things to do **86**

Bibliography **89**

A Selected R issues **92**
A.1 Numerical derivatives . . . 92
A.2 Estimating the Fisher information matrix . . . 93
A.3 Automated organization of models . . . 93
A.4 Doing leave-one-out cross-validation . . . 95
A.5 Executing the FIC analysis . . . 96
A.6 The simulated data . . . 98

B Experimental values for focus $F_{Y_{new}}^{-1}(0.01)$ and $F_{Y_{new}}^{-1}(0.95)$ **100**
B.1 Experimental results for $F_{Y_{new}}^{-1}(0.01)$ . . . 100
B.2 Experimental results for $F_{Y_{new}}^{-1}(0.95)$ . . . 100

C Selected theorems and proofs **104**
C.1 Asymptotic properties of the MLE . . . 104
C.1.1 Assumptions and regularity conditions . . . 104
C.1.2 Consistency of the MLE . . . 105
C.1.3 Asymptotic normality of the MLE . . . 106
C.2 Invariance of the MLE . . . 108
C.3 The delta method . . . 108
C.4 Elements of the FIC framework (FICology) . . . 108
C.4.1 The framework . . . 109
C.4.2 Limit distribution of parameters . . . 109
C.4.3 Limit distribution of a focus . . . 111

D List of abbreviations **113**

1 Introduction and main findings

Choosing the right model for estimation and prediction in the presence of an unknown data generating process is one of the core issues in statistical inference. A particular common application of model selection is choosing among covariates to include in regression models. For model selection, and in particular covariate selection, the statistician has a range of tools available. The traditional approach to this issue has been to perform various tests combined with stepwise elimination of variables. Such methods often leave substantial discretion to the statistician. However, more objective and pure data-based methods are available, such as the use of information criteria. Information criteria include the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the focused information criterion (FIC), and their use include the possibility of weighted averaging across candidate models.

Generally we can think that we use statistical inference to estimate and/or predict and that wrongful estimation/prediction is a source of loss. The better the model used, the less is the expected loss. Hence, a model has information-value. Several established information criteria, such as AIC and FIC, have an expected loss reduction interpretation.

Usually, model selection is based on the merits of the model as such. However, in regression settings where the measurement we wish to estimate or predict is dependent on gathering relevant covariates, there are types of applications where there are certain costs associated with getting hold of some or all of these covariate values. This may apply, for instance, to medical diagnostic, weather forecasts and financial forecasts. In these cases there can be a trade-off between using models providing highest information-value and the cost of gathering covariates to use in the estimation or prediction. This study aims at exploring principles for optimally deciding on the trade-off between information-value and costs in such settings. The study, hence, seeks to explore methods for a cost information-value trade-off in regression covariate selection. We delineate this study to generalized linear regression models (GLM), although the methods are also applicable in more general settings.

After establishing the role and place for this study in the context of the existing statistical literature, we will present a general framework for trading off cost and information-value in the GLM regression setting. This will be followed up by a more careful discussion on how covariate cost functions may look like and be constructed. After this we will present what can be considered to be a main part of this study, which is methods to estimate the expected loss associated with regression models for the purpose of finding the information-value of covariates. We will first look at prediction. We will see that in some situations we can analytically find approximately unbiased estimates directly by correcting the plug-in empirical distribution estimate for its bias, while in more general cases cross-validation is a preferred method. When it comes to estimation our approach will be to explore if FIC can be utilized for general loss functions. For this we will use Taylor-development as a general method. We will see that this method has some shortcomings when it comes to convergence. Hence, other methods should be explored where available. We will see that FIC works particularly well in combination with the LINEX loss function, which adds additional value to the FIC-framework. However, in general, when applying

FIC framework, we must be aware of the underlying assumptions of the FIC-framework, and take the appropriate precautions.

To keep this study within limits, we have been forced to make some hard priorities with respect to illustrating the methods on a data-set. Since we we want to illustrate different methods, differences between methods, and the performance of methods, we have found that it is most instructive to illustrate the methods on a simulated data set, where we have full control on the data generating process. The simulation experiment provide valuable insights. One valuable insight is that the methods produce fairly similar results as expected, indicating that the developed methods are indeed correct. When it comes to the applications based on FIC, we will see that those works fairly well as long as we take the necessary precautions related to the FIC-framework assumptions. In this context we will see that extra caution must be taken when applying the FIC-framework to models that are very wrong in terms of deviation from the true data generating process.

This study is of a practical nature. This means that we want to explore different methods and discuss advantages and disadvantages from a practical point of view. To stay within limits, this means that some theoretical details will have to be sacrificed. However, selected theorems with proofs are presented in Appendix C.

2 Statistical context and existing literature

There are several branches of overlapping literature relevant for this study. The theoretical basis for this study are the basic results in the main domain of what is usually covered by a graduate level statistical inference course. This includes topics such as GLM regression, estimation, bias-correction, bias-variance trade-off, and basic asymptotic theory. The main results can be found in graduate text books such as Casella and Berger (2001), Knight (2000), Wasserman (2003), and the more recent Boos and Stefansky (2013). We will go little further with respect to asymptotic theory, but not further than what can be found in introductory asymptotic theory textbooks, such as Polansky (2011).

Since the information value of models is crucial to this study, the literature on information criteria for statistical model selection naturally becomes highly relevant.[1] Information criteria are particularly relevant because they can be employed to say something about the information-value of a model or the relative information-value provided by alternative models. Information criteria stands in contrast to other techniques of model-testing, such as stepwise parameter hypothesis testing and F-tests, in providing a measurement that represents the information in a model.

A seminal contribution on the use information criteria to select among models is Akaike (1973), establishing the information criterion AIC ("An Information Criterion" as used by Akaike himself, but

[1] Statistical model selection can be informally be described as to use statistical methods and reasoning to choose a model based on a set of data. Statistical model-selection could be argued to be a subdicipline of model selection in more general. The choice between competing models and theories is a core issue in philosophy of science. Principles such as parsimony, in particular the use of "Occam's razor", has wider applications than just in statistics.

also well known as the "Akaike Information Criterion"). AIC is the value of the log likelihood-function inserted the maximum likelihood estimator (MLE) estimate subtracted the number of parameters.[2] AIC is based on an idea dating back to Boltzmann (1877) that a model can be seen as information, i.e. loss of entropy, about the true data generating process (DGP). These ideas were pursued further in, inter alia, contributions by Shannon (1948) and Kullback and Leibler (1951). Kullback and Leibler (1951) explored the concept that the information in a model can be seen as the distance, more precisely the Kullback-Leibler (K-L) distance[3], between knowing a model and knowing the DGP. Hence, there is a information loss from using a model, and the loss is the K-L distance to the true DGP. The seminal contribution of Akaike (1973) was to analyse the K-L distance in the context of traditional statistical measures such as log likelihood, MLE and Fisher information. AIC is in simple terms an estimate of the model specific part of the expected Kullback-Leibler distance between the model and the true data generating process (DGP) when using maximum likelihood estimators. Another way to say this is that AIC is an estimate of the expected relative K-L distance of a model when using MLE.

AIC served as a starting point for subsequent modifications of AIC, where many modifications impose different penalty terms on the number of parameters. One notable modification is TIC advocated by Takeuchi (1976), where the penalty for non-parsimony is the effective number of the parameters rather than the number of parameters. Another notable information criterion is BIC ("Bayesian Information Criterion") after Schwarz (1978). BIC is based on Bayesian reasoning, where, in simple terms, the model with highest probability is picked when BIC is used as a model selection criterion. Although the different theoretical foundation, BIC is as AIC (and TIC) based on the value of the log likelihood function, but with a different penalty term. BIC has the advantage over AIC that it asymptotically picks the correct model with probability one, while AIC does not.[4]

A more recent and innovative information criterion is FIC (Focused Information Criterion) introduced by Claeskens and Hjort (2003). This criterion is particularly innovative because it implements a conventional wisdom known to all modelers: which model is good depends on what you are using the model for. FIC is, in essence, a method for estimating the MSE (mean square error) of a given focus, for instance an upper quantile, for various candidate models. This can be used to choose the model with the lowest estimated MSE for the particular focus in question.

Several information criteria are based on a risk function, i.e the expected loss of errors, to be minimized. Hence, they can be interpreted as a tool to choose a model that minimizes expected loss when alternative models are available. In the application of AIC the loss can be seen as the K-L distance between the model and the true data generating process (DGP). FIC seeks to estimate the mean squared error (MSE) of a focus parameter for a given model. Thus, the squared error represent the loss in the risk function. This study follows the spirit of FIC in the sense that we assume that appropriate model

[2]There are different definitions on AIC, but limited to what multiplying factor to use.

[3]The K-L distance is not symmetric. Hence, the term "divergence" is often used instead. We will stay with the term "distance" in this study.

[4]See for instance Claeskens and Hjort (2008) p. 99 f.

selection is dependent on what we are using the model for.

The literature on information criteria is usually conserned with criteria for choosing the model with highest merits in a statistical sense. This is insufficient for our purpose as we want to take the costs of models into account. We are in a situation where we might be willing to sacrify some of the merits of a model to the benefit of a cheaper model. This means that we must have a loss function suitable for the specific decision context we are in to be traded off against costs. Since we both want to use context specific loss functions to calculate information value and we want to trade off information-value against models' cost, the literature on information criteria is not sufficient to address the topics in this study.[5]

The use of context specific loss functions in statistical decision making, and the reduction of this loss from gathering costly information, is not unfamiliar to statistics. In the literature on statistical decision theory the cost of gathering information is taken into account and must be economically traded off against value of gathering information in terms of reduced expected loss. This includes optimal sequential decsions solved by backward induction. Statistical decision theory experienced much development in the 1950s and onwards with the development of general decision theory.[6] Seminal contributions include Savage (1954), Raiffa and Schlaifer (1961), DeGroot (1970) and Berger (1985). From statistical decision theory, the theory of optimal sequential statistical decisions is particularly relevant for this thesis. This literature is Bayesian in nature. It is assumed that you experience a loss from making wrong decision. By gathering information you can update beliefs to make a more informed decision. Raiffa and Schlaifer (1961) showed that having more information always reduces expected loss (we will challenge this statement in Chapter 3). Thus, you should never say no to free information. The intuitive reason is that you reduce the amount of ex-ante uncertainty that never-the-less must be taken into account. However, as information is not free, there is a trade-off between the information-value gained from gathering more information and the cost of getting hold of this information. Very much in the spirit of this study, DeGroot (1984) speaks of changes in utility as the value of information. The value of information in this sense is how much obtaining it reduces your expected loss (or equivalently, increases your expected utility). In this literature, however, model uncertainty has traditionally not been taken into account. The model describing the probability of certain observations, i.e. the likelihood function, is assumed given. Hence, the DGP is assumed known. The Bayesian framework is, however, in principle suitable for incorporating model uncertainty. One could think of the model itself as a component of uncertainty and the value of the model is how much it changes expected loss. We simply add another level of uncertainty by assuming model uncertainty and allow for a priori probabilities for the models. This way of thinking about models, is of course not unfamiliar for Bayesians. Such kind model uncertainty is, in fact, the basis for the

[5] The necessity to sometimes include model costs in model selection is also aknowledged in Hjort and Claeskens (2003)

[6] A seminal contribution to decision theory was von Neumann and Morgenstern (1944). This contribution provided a analytical framework for the implementation of utility theory in decsion making. In particular relevant for the further development of statistical decision theory was the principle of expected utility maximization as equivalent to adhering to certain axioms considered as rational. Savage (1954) explored this framework further. While Von Neuman and Morgenstern was conserned with objective probabilities of different states, Savage estableshed the validity of expected utility maximization under subjective probabilities.

derivation of BIC. However, mixing the already existing literature on optimal sequential decisions with the theory of model uncertainty is to our knowledge not very well explored.[7] A probable reason for that is the practical computational complexity that soon appear in Bayesian analysis. A well known practical problem with Bayesian analysis is the calculation of complex a posteriori probabilities, which becomes even more complex when we incorporate model uncertainty. However, the analysis can be simplified by using the MLE as an approximation. In fact, the asymptotic properties of the MLE for Bayesian measurements, is central to the derivation of BIC. In this study we will not pursue the Bayesian perspective allowing for parameter and model probabilities. Rather we will use estimates of the expected loss associated with a model using data alone. Hence, Bayesian statistical decision theory will serve mostly as an inspiration and not as a theoretical framework for this study.

Computer algorithms can be used to perform automated searches for models that best fit the data. The literature on statistical learning provides algorithms for feature selection, which includes covariate selection, see for instance Witten et al. (2011) and Hastie et al. (2009). Statistical learning is in many ways a practical discipline where one uses the available methods at hand. However, cross-validation seems to be particularly popular in algorithmic feature selection. The reason is that cross-validation is easily applicable to most kinds of loss functions and most types of models, providing analytical desirable results without imposing much assumptions. Information criteria are also used for model selection in this literature where applicable.[8] In some cases cross-validation model selection corresponds with information criteria model selection.[9] However, algorithms taking into the account the cost of using a particular feature seems to not have reached the mainstream literature. The statistical learning literature provides some nice insight usable in this study. We will explore the usability of cross-validation methods in estimating the expected loss associated with a model. In this thesis, however, we will mainly assume that the number of candidate models are given and that the number of candidate models are not so big that we cannot compare the performance of all candidate models by brute force. Hence, the search-algorithm elements in the statistical learning literature, involving optimized algorithms to search for the best model among a large amount of candidate models, will not play a big role in this study.

3 Analytical framework and principles

3.1 Regression models

3.1.1 The structure of regression models

Real world data are generated by a true data generating process (DGP) that we usually don't fully know. Statistical inference is about making inferences about an unknown DGP. One way of doing so is to model

[7] However, Parmigani and Inoue (2009) p. 209 f. do indeed put up a general framework for the issue.

[8] See Hastie et al. (2009) Chapter 7.

[9] See for instance Claeskens and Hjort (2008) p. 51 f.

the DGP as a parametric regression model. In a regression model the distribution of a stochastic variable Y is dependent on several covariates $x_i, ..., x_p$ and a vector of parameters θ. We can write this as

$$Y_i \sim f(y \mid x_{i0}, ..., x_{ip}; \theta)$$

In a linear regression model the relationship between the response variable Y and the covariates can be expressed by a linear combination of the covariates. Thus, we can write

$$Y_i \sim f(y \mid \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma)$$

where we have that $\theta = \begin{bmatrix} \beta \\ \sigma \end{bmatrix}$. $\beta = (\beta_0, \beta_2, ..., \beta_p)^T$ is the vector of regression parameters linearly associated with the covariates, and $\sigma = (\sigma_o, ..., \sigma_q)^T$ are other parameters.

In this study we will assume that the covariates are non-stochastic variables, making the $Y_i's$ independent stochastic variables. This can be given several justifications. One justification might be that the covariates are genuinely non-stochastic, for instance, because the covariates are picked at will. Another interpretation, that will apply for this study, is that we will be interested in in-sample inferences, i.e. inferences based on the covariates in the sample. In other words, we will analytically be interested in the variations in the responses based on the covariates in the sample. A third justification, that is maybe not much of a justification, is that covariates are treated as non-stochastic in GLM regression, which we will return to just below.

From the very general form of regression described above, we can move to the normal classic linear regression, which is the most common form linear regression:

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \tag{3.1}$$

The residuals ε_i, i=1,...,n are independent and identically distributed (IID) following a normal distribution $N(0, \sigma^2)$. [10] This means that

$$Y_i \sim N(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma^2)$$

where the $Y_i's$ are independent since we are assuming that the regression variables are non-stochastic.

The normal linear regression can be considered as special case of generalized linear models (GLM). GLM are characterized by the following distribution:

$$Y_i \sim f(y_i; \theta_i, \varphi)$$

[10] Note that classic linear regression is often introduced without making assumptions regarding the distribution of the residuals. Rather the residuals are assumed to be IID, with expectation zero and constant variance. If we use the method of least squares to estimate the parameters, we don't need to know the distribution of the residuals, we only need to make some less restrictive assumptions. However, if we use the method of maximum likelihood, distributional assumptions are needed. We will return to estimation issues just below.

where Y_i are independent and $f(y;\theta_i,\varphi)$ almost belongs to the exponential family (also known as overdispersed or generalized exponential family). In the univariate case, we can write the log of the density as

$$\log f(y_i;\theta_i,\varphi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\varphi)} + c(y_i,\varphi)$$

The dispersion term $a_i(\varphi)$ separates the distribution from a pure exponential family. Under this specification, we have $\mu_i = E(Y_i) = b'(\theta_i)$ and $VAR(Y_i) = b'(\theta_i)a_i(\varphi)$.[11] GLM includes many well known distributions, including the normal, Poisson and binomial. As mentioned above the normal distribution is probably the mostly used. The Poisson distribution is suitable in the estimation of count data, for instance the number of cars that passes a point within a time interval. Binomial regression is suitable in the estimation of probabilities. The response Y can then take the value 0 or 1, and the purpose of the regression is then to estimate the probability that Y equals 1, given some covariates.

The covariates enters the distribution with $\eta_i = \beta + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ in the way that the mean μ_i is a smooth and invertible function of η_i. That means that we can write

$$\mu_i = m(\eta_i)$$

This gives a link function g,

$$\eta_i = m^{-1}(\mu_i) = g(\mu_i)$$

The parameter θ_i is referred to as the natural or canonical parameter. Often we have the canonical link $\theta_i = g(\mu_i)$. There can be several link function within a class of GLM. For instance is the logit, given by $\eta_i = \log(\frac{\mu_i}{1-\mu_i})$, one of several link functions used in the binomial regression model. In the Poisson rgeression model one usually uses $\eta_i = \log(\mu_i)$. These are both canonical links.

As mentioned above we will mainly delineate this study to GLM regression models. The reason for that is two-folded. Firstly, we can assume independent $Y_i's$. Secondly, the exponential class satisfies regularity conditions enabling us to rely on general statistical results satisfied under these regularity conditions.[12]

3.1.2 Estimation and prediction

Since the DGP is unknown, the parameters of our regression model are also naturally unknown. A crucial part of statistics is to estimate the regression parameters and make inferences about the regression parameters based on observed data. There are several methods that can be employed to estimate parameters. The most employed methods are the method of moments, the method of least squares and the method of maximum likelihood. Often the outcome of these methods coincide. In this study, we will use the method

[11] See Boos and Stefansky (2013) section 2.3.3 for a brief, but precise description of GLM. See also McCullagh and Nelder (1989) for a complete description of generalized linear models.

[12] For some theoretical details on regularity conditions, see Appendix C.

of maximum likelihood as the basis for parameter estimation. The reason for this is that maximum likelihood is a general method with nice statistical properties, as will be described in the next subsection. For now, let us assume that the parameter estimators are $\hat{\theta}_n$, where n is the number of observations used for estimation.[13]

Instead, or in addition, to making inferences on θ, we may want to make inferences on a function of the parameters, $\mu(\theta)$. For instance we might be interested on making inferences on the expectation of Y, given some particular combination of covariates x_0, i.e $E(Y \mid x_0;\theta)$. We might also, for instance, be interested in the probability that Y is less than some particular value α for some particular combination of covariates x_0, i.e. $P(Y < \alpha \mid x_0;\theta)$. An obvious candidate for this estimator is the well know plug-in estimator $\widehat{\mu}_n(\theta) = \mu(\hat{\theta}_n)$, obtained by replacing the parameters with the MLE in the function.

In addition to making inferences about parameters, including functions of parameters, we are often interested in prediction. In a prediction setting there is a so-called irreducible uncertainty. This can best be explained by inspecting the classical normal linear regression in equation (3.1) above. Assume that we are going to predict Y_0 from a new combination of covariates, x_0. No matter how good we are in estimating the parameters $\hat{\theta}_n = \begin{bmatrix} \hat{\beta}_n \\ \hat{\sigma}_n \end{bmatrix}$, we will still be left with the uncertainty ε_0 when we try to predict Y_0. This will always leave us with an irreducible error in prediction.

In this study we will be concerned with both estimation and prediction. As we will see, different methods of model selection can be appropriate in calculating the information value of a model, dependent on whether our concern is estimation or prediction.

3.2 Maximum likelihood estimators

3.2.1 Maximum likelihood estimators and their properties

The maximum likelihood estimators (MLE) are of crucial interest to this study as our analysis will be based on the MLE. We will first illustrate the main points and properties of MLE by assuming IID variables, and then return to the regression setting below.

The MLE's are found by maximizing the likelihood function, i.e

$$\hat{\theta}_n = \arg\max \mathscr{L}_n(\theta)$$

where $\mathscr{L}_n(\theta) = \prod_{i=1}^{n} f(y_i;\theta)$ is the likelihood function for IID variables. Hence, we choose parameter estimates that maximize the "likelihood" of the data. Usually, we instead operate with (and maximize) the log likelihood function

$$\ell_n(\theta) = \log \mathscr{L}_n(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$$

[13] We will generally use the number of observations as subscript to estimators.

We will see just below that operating with the log-likelihood has much more reasons than just that is computationally easier to work with than the likelihood itself. The maximum likelihood estimators $\hat{\theta}_n$, are usually uniquely found by solving $\ell_n'(\hat{\theta}_n) = 0$. However, in some cases the maximum might not be an interior soulution. Since we in this study will assume that the models are within the almost exponential family of the GLM framework, the MLEs will be unique and interior.

Under certain regularity conditions, the MLE have many nice properties.[14] For simplicity, we will first assume that the model we are estimating is the "true" model, i.e. the actual DGP. In other words, we assume that the model we estimate, $f(y;\theta)$ corresponds to the DGP, $g(y)$, for Y, for a certain θ.

In this case, the MLE is consistent, i.e $\hat{\theta}_n \xrightarrow{p} \theta$. Furthermore, the MLE is asymptotically normal. Let $s(y;\theta) = \partial log f(y_i;\theta)/\partial\theta$ be the score function and $I(y;\theta) = \partial^2 log f(y_i;\theta)/\partial\theta\partial\theta^t$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, J^{-1}KJ^{-1}) \tag{3.2}$$

where $K = E[s(Y;\theta)s(Y;\theta)^t] = VAR[s(Y;\theta)]$ is the Fisher information matrix[15], which equals $J = -E[I(Y;\theta)]$, when $f(y;\theta)$ corresponds to the DGP, $g(y)$.[16] Since $J = K$, Equation (3.2) reduces to

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, J^{-1}) \tag{3.3}$$

The proof of the consistency of the MLE and the limiting normal distribution of MLE given in equation (3.2) can be derived under more general circumstances than the case where $f(y;\theta)$ corresponds to the DGP, $g(y)$, and relies on the concept of Kullback-Leibler distance which we will return to now.

The MLE has nice properties even if the model to be estimated, $f(y;\theta)$, not necessarily corresponds with the actual DGP, $g(y)$. We have a "parallel" to the consistency of MLE for this situation. When $f(y;\theta)$ does not correspond to the actual DGP, $g(y)$, we can talk about the least false parameters. This requires some further explanation. The Kullback-Leibler (K-L) distance[17] between a model and a true DGP is given by

[14] A summary of the properties of the MLE estimators can be found in Wasserman (2003) p. 122 et seq. See also Knight (2000) Chapter 5. The proofs are included in most intermediate textbooks on statistical inference. We will discuss some of the most important properties and indicate the idea behind the proofs. More details on selected proofs are provided in Appendix C.

[15] The term information matrix is explained by its role in determining the Cramér-Rao lower bound for any estimator, see Casella and Berger (2001) p. 335.

[16] This holds "in all but rare cases", see Knight (2000) p. 265. It certainly hold for the exponential family used for GLM, which we use in this study. A proof is provided in Appendix C.

[17] As mentioned in Chapter 2, the distance is often referred to as divergence due to lack of symmetry, and one can sometimes see the notation $D_{KL}(g \parallel f)$. We will use the terme "distance" and use a slightly easier notation.

$$\begin{aligned} D(g(y), f(y;\theta)) &= E[\log\frac{g(y)}{f(y;\theta)}] \\ &= \int \log\frac{g(y)}{f(y;\theta)} g(y)dy \\ &= \int \log(g(y))g(y)dy - \int \log(f(y;\theta))g(y)dy \\ &= E[\log(g(y)] - E[\log(f(y;\theta)] \end{aligned}$$

The K-L distance can be interpreted as the loss of information by relying on the information in the model rather than knowing the full DGP. The shorter the distance between the model and the DGP, the better. The lower bound if the distance is 0, which it would be only if $f = g$ almost everywhere.[18] Let θ_0 be the parameters that minimize the K-L distance from $f(y;\theta)$ to g. Hence,

$$\theta_0 = \arg\min D(g(y), f(y;\theta))$$

θ_0 can be said to be the least false parameters. We have that $\hat{\theta}_n \xrightarrow{p} \theta_0$.[19] In fact, under mild regularity conditions, we have strong consistence, i.e $\hat{\theta}_n \xrightarrow{a.s} \theta_0$. Hence, the MLE estimators are consistent estimators for the least false parameters achieved by minimizing the K-L distance. A special case is if $f(y;\theta)$ equals $g(y)$, where θ_0 is the parameters of the true DGP as described above. The asymptotic normality property in still hold, but now with

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, J^{-1}KJ^{-1}) \tag{3.4}$$

Note that the J and K are calculated with respect to the true DGP, i.e. that $J = -E_g I(Y;\theta_0)$ and $K = E_g[s(Y;\theta_0)s(Y;\theta_0)^t] = VAR_g[s(Y;\theta_0)]$ where the subscript g is meant to clarify that the expectation and variance are calculated with respect to g. The proof of equation (3.4) is based on approximating $\ell_n'(\hat{\theta}_n)$ by a first order Taylor development of $\ell_n'(\theta)$ around θ_0, and using that $\ell_n'(\hat{\theta}_n) = 0$ and asymptotic theory.[20] Also note that when $f(y;\theta)$ does not necessarily correspond with the actual DGP, $g(y)$, we cannot generally say that $K = J$ as above. $J^{-1}KJ^{-1}$ is known as the sandwich matrix in the statistical literature. Note that when $f(y;\theta) = g(y)$, such that $J = K$, then $J^{-1}KJ^{-1} = J^{-1}$ and we are back to the simpler formula above.

In an estimation setting we do not know J and K since we don't know the DGP. However, we can get asymptotically good estimates by using the MLEs as a substitute to the actual parameters and take the average over the sample. Hence, we have approximately that in the limit

[18] Claeskens and Hjort (2008) p. 66

[19] For a proof see Knight (2000) p. 260 and Wasserman (2003) p. 126. See also Casella and Berger (2001) for a slightly different type of proof. A sketch of proof is privided in Appendix C.

[20] See Knight (2000) p. 263. A sketch of proof is provided in Appendix C.

$$\hat{\theta} \sim N(\theta_0, \frac{1}{n}\hat{J}_n^{-1}\hat{K}_n\hat{J}_n^{-1})$$

which means that

$$VAR(\hat{\theta}_n) \approx \frac{1}{n}\hat{J}_n^{-1}\hat{K}_n\hat{J}_n^{-1}$$

Where

$$\begin{aligned}\hat{J}_n &= -\frac{1}{n}\frac{\partial \ell_n(\theta)}{\partial\theta\partial\theta^t}\bigg|_{\theta=\hat{\theta}_n} = -\frac{1}{n}\sum_{i=1}^{n} I(y_i;\hat{\theta}_n)\\ \hat{K}_n &= \frac{1}{n}\sum_{i=1}^{n} s(y_i;\hat{\theta}_n)s(y_i;\hat{\theta}_n)^t\end{aligned}$$

If our model $f(y;\theta)$ is reasonably close to the true DGP, $g(y)$, then we should have $\hat{J}_n^{-1} \approx \hat{K}_n$ and thus

$$VAR(\hat{\theta}_n) \approx \frac{1}{n}\hat{J}_n^{-1}$$

The MLE has additional useful properties. The MLE is invariant, i.e $\mu(\hat{\theta}_n)$ is the MLE of $\mu(\theta)$ for functions μ.[21] The MLE is also asymptotically efficient. More precisely, the variance of the MLE converges towards the Cramér-Rao lower bound.[22] Finally, the MLE is also approximately the Bayes estimator.[23] We will briefly return to the Bayes estimator below.

In this study we will be concerned with regression models within the GLM framework. The observations will still be assumed to be independent, but they will not be identically distributed, as the distribution of each obseravtion will be assumed to be dependent on the value of the covariates. The MLE likelihood framework and the properties of MLE are easily expanded to such regression framework.[24] The log likelihood function now becomes

$$\ell_n(\theta) = \log \mathscr{L}_n(\theta) = \sum_{i=1}^{n} \log f(y_i \mid x_{i0},...,x_{ip};\theta))$$

We can now write

$$\begin{aligned}J_n &= \frac{1}{n}\sum_{i=1}^{n} -E_g I(Y \mid x_i;\theta_{0,n}) \qquad (3.5)\\ K_n &= \frac{1}{n}\sum_{i=1}^{n} E_g[s(Y \mid x_i;\theta_{0,n})s(Y \mid x_i;\theta_{0,n})^t]\end{aligned}$$

[21] For a proof, see Casella and Berger (2001) p. 320. A sketch of proof is provided in Appendix C.
[22] For a proof, see Casella and Berger (2001) p. 472.
[23] Wasserman (2003) p. 126.
[24] See Claeskens and Hjort (2008) p. 27.

as the parallells to J and K under the empirical distribution of the covariates, C_n. We now have the true DGP, $g(y \mid x)$, and $\theta_{0,n}$ is the least false parameters according to the distribution of covariates, C_n. When taking the assumption of non-stochastic covaraites seriously, this distribution is more naturally to be considered as weights. In the limit we have under natural conditions:[25]

$$J_n \xrightarrow{p} J$$
$$K_n \xrightarrow{p} K$$

for some limits J and K. We then have

$$\sqrt{n}(\hat{\theta}_n - \theta_{0,n}) \xrightarrow{d} N(0, J^{-1}KJ^{-1}) \tag{3.6}$$

Estimators for J_n and K_n are

$$\begin{aligned} \hat{J}_n &= -\frac{1}{n}\sum_{i=1}^{n} I(y_i \mid x_i; \hat{\theta}_n) \\ \hat{K}_n &= \frac{1}{n}\sum_{i=1}^{n} s(y_i \mid x_i; \hat{\theta}_n) s(y_i \mid x_i; \hat{\theta}_n)^t \end{aligned}$$

3.2.2 The delta method

A nice complement to maximum likelihood estimation is the delta-method.[26] Let $\mu(\theta)$ be a function with continous partial derivatives of the model parameters and that satisfies some additional technical assumptions we will not delve into here. By using first order Taylor developments, and as a result of the limiting normal distribution of $\hat{\theta}_n$ explained in section 3.2.1, we have

$$\sqrt{n}(\mu(\hat{\theta}_n) - \mu(\theta_0)) \xrightarrow{d} N(0, (\nabla\mu)^t J^{-1} K J^{-1} (\nabla\mu))$$

where $\nabla\mu = (\frac{\partial\mu}{\partial\theta_1}, ..., \frac{\partial\mu}{\partial\theta_k})^t$ is the gradient vector of μ with respect to the parameters. This is an application of the delta-method or delta-theorem. If $f(y;\theta)$ corresponds to the true DGP, then the limit distribution reduces to

$$\sqrt{n}(\mu(\hat{\theta}_n) - \mu(\theta_0)) \xrightarrow{d} N(0, (\nabla\mu)^t J^{-1} (\nabla\mu))$$

Since we dont know the parameter values of $\nabla\mu = (\frac{\partial\mu}{\partial\theta_1}, ..., \frac{\partial\mu}{\partial\theta_k})^t$, we use the MLE as plug-in. Because of the consistency of MLE, this works well in the limit.

[25] See Claeskens and Hjort (2008) p. 27.

[26] See Casella and Berger (2001) p. 242 or Knight (2000) p. 130 for proof (at least for the univariate case) and details. A general version of the theorem with a sketch of proof is provided in Appendix C. See also Wasserman (2003) p. 131 for a good intuitive description.

3.3 Covariates and information value

Models provide more or less information about the true nature of a data generating process (DGP). If we use a model M that includes a subset of possible covariates, we get more or less information on the DGP, depending on how much the covariates we have included provide information about the DGP.

We will use the notation that M_k denotes model k, which is one of m candidate models, i.e. k=1,..,m. $M_{\{abc\}}$ corresponds to the the model where the covariates number a,b and c are included. For instance $M_{\{012\}}$ corresponds to the model $Y \sim f_Y(y \mid x_0, x_1, x_2; \theta)$ and $M_{\{045\}}$ corresponds to the model $Y \sim f_Y(y \mid x_0, x_4, x_5; \theta)$. Hence if our two (only) candidate models are $M_{\{012\}}$ and $M_{\{045\}}$, we have two candidate models $M_1 = M_{\{012\}}$ and $M_2 = M_{\{045\}}$.

To isolate the problem of covariate selection, we will not mix categories of regression models in covariate selection. Hence, if we for instance operate with a normal linear regression model, we will assess the information-value of covariates within this model category. Hence, we will not create combinations of covariates and model types (i.e. different GLM categories). To compare both covariate combinations and model category combinations would, however, be a possible extension of this study.

To assess how much information a model provides, we must have some way to measure this information value. We will first look at two commonly used measurements of model information; the expected Kullback-Leibler (K-L) distance and the mean squared error. These can be considered as special cases of expected loss. Reduction in expected loss can be seen as a general way to measure information value, which will be explored next. We will then finalize this subsection with some terminology discussion and Bayesian perspectives.

3.3.1 Reduced Kullback-Leibler distance as information value

To illustrate reduced Kullback-Leibler distance as information value we will first use the notation $f(y; \theta_{M_k})$ as the density for model M_k, ignoring the covariates $x_0, ..., x_p$. The reason for this is to not let the complexity of including the covariates complicate the principles we want to highlight. Hence we will introduce the concept of reduced Kullback-Leibler distance as information value by assuming IID, and return to the inclusion of covariates below after the concept is introduced.

The K-L distance for model M_k to the true DGP, $g(y)$, is given by

$$\begin{aligned} D(g(y), f(y;\theta_{M_k})) &= E(\log \frac{g(y)}{f(y;\theta_{M_k})}) \\ &= \int \log(g(y))g(y)dy - \int \log(f(y;\theta_{M_k}))g(y)dy \\ &= E[\log(g(y))] - E[\log(f(y;\theta_{M_k}))] \\ &= C - R(M_k) \end{aligned}$$

C is a constant independent of the model and $R(M_k)$ is a model specific term. As explained above, the lower bound of the K-L distance is zero, which it would only be if $f(y;\theta_{M_k}) = g(y)$ almost everywhere.

However since we don't know θ_{M_k}, we instead rely on the MLE, $\hat{\theta}_{n,M_k}$. Recall, that MLE is a consistent estimate for θ_{M_k} that minimizes the K-L distance as explained above. Hence, to derive the information value, we will operate with the expected K-L distance, where the expectation is taken over the MLE as a stochastic variable. This gives

$$\begin{aligned} E[D(g(y), f(y;\hat{\theta}_{n,M_k}))] &= E[E(\log\frac{g(y)}{f(y;\hat{\theta}_{n,M_k})})] \\ &= \int \log(g(y))g(y)dy - E[E(\log(f(y;\hat{\theta}_{n,M_k})] \\ &= \int \log(g(y))g(y)dy - E[\int \log(f(y;\hat{\theta}_{n,M_k}))g(y)dy] \\ &= C - E(R(M_k)) \\ &= C - Q(M_k) \end{aligned}$$

Hence, we have that

$$\begin{aligned} Q(M_k) &= E[R(M_k)] \\ &= E[E(\log(f(y;\hat{\theta}_{n,M_k})] \end{aligned}$$

varies with the models. The outer expectation is with respect to the MLE.

The empirical distribution plug-in estimate of $Q(M_k)$ is

$$Q_n^*(M_k) = \frac{1}{n}\sum_{i=1}^{n}\log(f(y_i;\hat{\theta}_{n,M_k}))$$

By the law of large numbers, $Q_n^*(M_k)$ is a consistent estimator of $Q(M_k)$. However, it is not unbiased. Under certain conditions, which here can be heuristically summarized as assuming that the estimated model is not too far from the true data generating process[27], we have that

$$E(Q_n^*(M_k)) \approx Q(M_k) + \frac{|M_k|}{n} \tag{3.7}$$

where $|M_k|$ is the number of estimated parameters (dimension) for M_k. A way to explain this is that $Q_n^*(M_k)$is biased upwards because of the "sample bias" following from that we re-use the data from the parameter estimation process. The proof of equation (3.7) is based on a second order Taylor development of both $Q_n^*(M_k)$ and $R(M_k)$ giving us a limiting distribution for $Q_n^*(M_k) - R(M_k)$. This shows that expected $Q_n^*(M_k)$ "overshoots" the target $Q(M_k)$ by $\frac{Tr(J^{-1}K)}{n}$, where J and K are defined in the discussion of the properties of the MLE above.[28] When the candidate model is the true DGP, $g(y)$, we know that $J = K$,

[27] In the terms used above, this means that J is not too far from K. This we be explained in more detail below.

[28] See Claeskens and Hjort (2008) p. 31.

and hence $Tr(J^{-1}K) = Tr(I_{|M_k|}) = |M_k|$. If the candidate model is not "too far" from the DGP, we have $Tr(J^{-1}K) \approx |M_k|$. However, it would be more correct to use $Tr(J^{-1}K)$ in the bias correction term. This term can be considered as the effective number of parameters.[29] Since we don't know J and K, we could use $Tr(\hat{J}_n^{-1}\hat{K}_n)$. This is in essence what is used in the TIC modification of AIC. We will return to AIC just below. We are not sure, however, that using $Tr(\hat{J}_n^{-1}\hat{K}_n)$ would be superior to using $|M_k|$, at least for a low sample size.[30]

Hence, if we use

$$\widehat{Q}_n(M_k) = Q_n^*(M_k) - \frac{|M_k|}{n} = \frac{1}{n}\sum_{i=1}^{n}\log(f(y_i;\hat{\theta}_{n,M_k})) - \frac{|M_k|}{n}$$

as the estimator of $Q_n(M_k)$, then we will have an approximately unbiased estimate. Note that $\widehat{Q}_n(M_k)$ is not an estimate for the expected K-L distance. For that, we also need an estimate for $\int \log(g(y))g(y)dy$.

The information criterion AIC is based on using $\widehat{Q}_n(M_k)$ to compare models. AIC is not unanimously defined, but is always $\widehat{Q}_n(M_k)$ multiplied by various constants. In a majority of the literature, $\widehat{Q}_n(M_k)$ is multiplied by n to get rid of n in the definition of AIC. Furthermore, $\widehat{Q}_n(M_k)$ is often multiplied by 2 for historical reasons. Hence, AIC is usually obtained by multiplying $\widehat{Q}_n(M_k)$ by 2n. We then get

$$AIC(M_k) = 2n\widehat{Q}_n(M_k) = 2\sum_{i=1}^{n}\log(f_Y(y_i;\hat{\theta}_{M_k})) - 2|M_k| = 2\ell_n(\hat{\theta}_{n,M_k}) - 2|M_k| \tag{3.8}$$

where $\ell_n(\hat{\theta}_{n,M_k}) = \sum_{i=1}^{n} log(f(y_i;\hat{\theta}_{n,M_k}))$. Under this definition, the larger AIC is better, since the interpretation would be that this reduces the expected K-L distance to the true DGP.

As for the MLE framework part, the expected K-L distance estimation leading to AIC can quite easily be expanded to the regression framework.[31] In the regression setting we let

$$Q_n(M_k) = E[\frac{1}{n}\sum_{i=1}^{n}E(\log(f(y \mid x_i;\hat{\theta}_{n,M_k})]$$

By analogy of the IID case, we get that an approximately unbiased estimator for $Q_n(M_k)$ is

$$\widehat{Q}_n(M_k) = \frac{1}{n}\sum_{i=1}^{n}\log(f(y_i \mid x_i;\hat{\theta}_{M_k})) - \frac{|M_k|}{n}$$

leading us to the same AIC formula as for the IID case.

[29] See Hastie et al. (2009) p. 232.

[30] Burnham and Anderson (2002) p. 66 points out that using $Tr(\hat{J}_n^{-1}\hat{K}_n)$ requires the estimation of many measurements, and warn against its use.

[31] See Claeskens and Hjort (2008) p. 31.

We can now estimate the value of information in terms of reduced expected K-L distance of adding more parameters, for instance by adding an additional covariate in a regression model setting (assuming that adding a covariate means to add another parameter). Assume we have two competing models $f(y;\theta_{M_i})$ and $f(y;\theta_{M_j})$. Then the reduction in expected K-L distance using model j relative to model i is

$$E[D(g,f(y;\hat{\theta}_{n,M_i}))]-E[D(g(y),f(y_i;\hat{\theta}_{n,M_j}))]=-Q_n(M_i)+Q_n(M_j)=Q_n(M_j)-Q_n(M_i)$$

By using $\widehat{Q}(M_k)$as an estimator we get

$$\widehat{E}_n[D(g,f(y;\hat{\theta}_{n,M_i}))]-\widehat{E}_n[D(g(y),f_Y(y;\hat{\theta}_{n,M_j}))]=\frac{1}{2n}(AIC(M_j)-AIC(M_i)) \tag{3.9}$$

Equation (3.9) gives an estimate of the information-value of a better model, assuming the value of information is reduced expected K-L distance to the true DGP. Assume that model j is obtained by adding a covariate to model i. Then $AIC(M_j)-AIC(M_i)=2(\ell_n(\hat{\theta}_{n,M_j})-\ell_n(\hat{\theta}_{n,M_i})-1)$. Hence, the log likelihood must increase by at least one for AIC to improve. The estimated reduction in K-L distance is $\frac{\ell_n(\hat{\theta}_{n,M_j})-\ell_n(\hat{\theta}_{n,M_i})-1}{n}$. This will be the estimated information-value of the additional covariate if the the value of information is the expected reduced K-L distance.

3.3.2 Reduced mean squared error as information value

Minimizing mean squared error (MSE) has strong traditions in statistical analysis. For instance, MSE is used to derive the best linear predictor in forecasting, and minimizing the empirical MSE lies behind the least squared method of estimation. Furthermore, and simplified, one can say that minimizing MSE lies behind traditional model selection techniques, such as R^2-evaluation and some F-tests.

MSE is in particular instructive because of its easy decomposition into variance and bias squared in parameter estimation. We have in general that for a parameter estimate $\hat{\theta}_n$:

$$\begin{aligned} MSE(\hat{\theta}_n) &= E[(\hat{\theta}_n-\theta)^2] \\ &= E[(\hat{\theta}_n-E(\hat{\theta}_n))^2]+(E(\hat{\theta}_n)-\theta)^2 \\ &= VAR(\hat{\theta}_n)+BIAS^2(\hat{\theta}_n) \end{aligned} \tag{3.10}$$

Hence, the MSE of the estimator can be decomposed into a trade-off crucial to statistical analysis: the bias-variance trade-off. The bias-variance trade-off also appear in prediction settings. Assume that we want to generally predict $Y=g(x)+\varepsilon$ by some estimated regression function $\widehat{g}_n(x)$, where x is a vector

of covariates and $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. We can then write

$$\begin{aligned} MSE(\widehat{g}_n(x)) &= E[(Y - \widehat{g}_n(x))^2] \\ &= E[(Y - E(Y \mid x))^2] + E[(E(Y \mid x) - \widehat{g}_n(x))^2] \\ &= \sigma^2 + E[(E(Y \mid x) - \widehat{g}_n(x))^2] \\ &= \sigma^2 + E[(\widehat{g}_n(x) - E(\widehat{g}_n(x)))^2] + (E(\widehat{g}_n(x)) - g(x))^2 \\ &= \sigma^2 + VAR(\widehat{g}_n(x)) + BIAS^2(\widehat{g}_n(x)) \end{aligned} \tag{3.11}$$

As we see from equation (3.11), there is a irreducible uncertainty σ^2 when we do prediction because of the ε term in addition to the bias variance trade-off. No matter how good our prediction is, we cannot avoid the uncertainty associated with ε in prediction. This example is also illustrative for the difference between estimation and prediction. In estimation, the irreducible uncertainty is not present. As a consequence the estimator might converge in probability towards its true value, while this is not possible for prediction. A third result of equation (3.11) is that if we want to make a predictor that minimize MSE in the prediction context above, we might choose the predictor that minimize $E[(E(Y \mid x) - \widehat{g}_n(x))^2]$ since we cannot do anything about σ^2 anyway.

The FIC (Focused Information Criterion) framework develops the minimization of a focused MSE into an information criterion for model selection (and model averaging). The approach taken by FIC is that the loss can be considered as the mean square error of some focus in question for a candidate model. A focus in the FIC sense is a function of parameters, i.e. a parameter. A focus can typically be to estimate the expected value in a regression model setting, but the focus can be whatever we are interested in. In financial analysis, we might of course be interested in the expected payoff of a financial portfolio, but for risk management we might also be interested in estimating the 5 percent percentile of the payoff. With FIC we aim to choose the best model for estimating the focus in question. When using FIC for model selection, the model associated with the lowest FIC should be chosen, as this the model which gives the lowest estimated MSE for a given focus. In other words, the value of using a better model is the reduced MSE for a given focus.

A main innovation by FIC is the use of limit distributions to obtain an estimate of the MSE of a focus associated with a particular model. To motivate FIC and highlight the main principles, we will, as above, start with the IID situation and explain how the principles more or less easily can be expanded to the regression situation.

In the derivation of FIC, it is assumed that there is a wide true model. The wide model entails all possible narrower models. Hence, we have system of nested models. The general idea is to check if some narrower model entailed by the wide model is better in minimizing the MSE for a focus parameter μ estimated by MLE. The MSE is estimated by employing asymptotic theory for the limit distribution of the focus parameter. FIC is based on limit results, and for a given n, the results are approximate.

In the FIC framework it is assumed that Y has the density

$$f_n(y) = f(y; \theta_0, \gamma_0 + \delta/\sqrt{n}) \tag{3.12}$$

θ_0 is a vector of those parameters that are always included (protected parameters) and is assumed to be of of dimension p. $\gamma = \gamma_0 + \delta/\sqrt{n}$ represent the free parameters, which is of dimension q. The subset of models $\{M_i\}_{i=1,\ldots,m}$ compromise the various models between the full (wide) model and the narrowest model.

An explanation for this particular model construction is due. Why operate with a model that appears to violate Kolmogorovs consistency assumptions by letting the data generating process beeing dependent on the number of data? The reason for this is that we get the variance and bias squared on the same $\frac{1}{n}$-scale as will become apparent below. The variance and bias-squared become exchangeable currencies, as Hjort and Claeskens (2003) elegantly put it.

Let $\mu_{true} = \mu(\theta, \gamma)$ be the true value of the focus, and $\hat{\mu}_{n,M_i}$ the estimated focus under model M_i obtained by plugging in the MLE estimates obtained under model M_i. Hence, $\hat{\mu}_{n,M_i}$ is the MLE under model M_i. Let J be the Fisher information matrix for the wide model, which can be split into the protected and free parameters. Hence,

$$J = \begin{bmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{bmatrix} \text{ and } J^{-1} = \begin{bmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{bmatrix}$$

By using asymptotic theory, first order Taylor expansions, delta method principles and straight forward matrix manipulations [32], we get that for the MLE of δ in the wide model, we have

$$\hat{\delta}_{n,wide} = \sqrt{n}(\hat{\gamma}_{n,wide} - \gamma_0) \xrightarrow{d} D \sim N_q(\delta, Q) \tag{3.13}$$

and

$$\sqrt{n}(\hat{\mu}_{n,M_i} - \mu_{true}) \xrightarrow{d} \Lambda_{M_i} = \Lambda_0 + \omega^t(\delta - G_{M_i}D) \tag{3.14}$$

where

$$\begin{aligned} \Lambda_0 &\sim N(0, \tau_0^2) \\ Q &= J^{11} \\ \tau_0^2 &= \left(\frac{\partial \mu}{\partial \theta}\right)^t J_{00}^{-1} \frac{\partial \mu}{\partial \theta} \\ \omega &= J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma} \\ G_{M_i} &= Q_{M_i}^0 Q^{-1} \\ Q_{M_i}^0 &= \pi_{M_i}^t Q_{M_i}^{-1} \pi_{M_i} \\ Q_{M_i} &= (\pi_{M_i} Q^{-1} \pi_{M_i}^t)^{-1} \end{aligned}$$

[32] See Theorem 6.1 in Claeskens and Hjort (2008) with a sketch of the proof. The full proof can be found in Claeskens and Hjort (2003). The principles behind the proof is also presented in Appendix C.

π_{M_i} is the appropriate projection matrix of dimension $(|M_i| - p) \times q$ of zeros and ones. For instance, if q=2 and we only include the second of the free parameters in model M_i, then $\pi_{M_i} = \begin{bmatrix} 0 & 1 \end{bmatrix}$. Consequently we have the following limits based on the limit distribution

$$nVAR(\hat{\mu}_{n,M_i}) = (\tau_0^2 + \omega^t Q_{M_i}^0 \omega)$$

which we also can write as

$$VAR(\hat{\mu}_{n,M_i}) = \frac{1}{n}(\tau_0^2 + \omega^t Q_{M_i}^0 \omega)$$

and

$$\begin{aligned} \sqrt{n}BIAS(\hat{\mu}_{n,M_i}) &= E(\sqrt{n}(\hat{\mu}_{n,M_i} - \mu_{true})) \\ &= (\omega^t(\delta - G_{M_i}\delta)) \\ &= (\omega^t(I_q - G_{M_i})\delta) \end{aligned}$$

which also can be written as

$$\begin{aligned} BIAS(\hat{\mu}_{n,M_i}) &= \frac{1}{\sqrt{n}}(\omega^t(\delta - G_{M_i}\delta)) \\ &= \frac{1}{\sqrt{n}}(\omega^t(I_q - G_{M_i})\delta) \end{aligned}$$

A rather counter-intuitive observation is that the bias seem to approach zero as n grows. However, this is not surprising by inspecting the density of the assumed DGP in equation (3.12) if we assume we have got γ_0 right. However if γ_0 is set to some other value in the estimation process, for instance 0, the bias willl not approach zero since δ grows at the same rate, $\sqrt{n}$, for a given true underlying fixed parameter value.

Using the asymptotic result above, we find that $\hat{\mu}_{n,M_i}$ has the following limiting mean squared error

$$\begin{aligned} MSE(\hat{\mu}_{n,M_i}) &= E[(\hat{\mu}_{n,M_i} - \mu_{true})^2] \\ &= Var(\hat{\mu}_{n,M_i}) + (E(\hat{\mu}_{n,M_i}) - \mu_{true})^2 \\ &= VAR(\hat{\mu}_{n,M_i}) + BIAS^2(\hat{\mu}_{n,M_i}) \\ &= \frac{1}{n}(\tau_0^2 + \omega^t Q_{M_i}^0 \omega + \omega^t (I_q - G_{M_i})\delta\delta^t (I_q - G_{M_i})^t \omega) \end{aligned} \tag{3.15}$$

The first two terms are the variance part and the last term is the bias squared part. I_q is the identity matrix of dimension q. Note that for the full model, we have $G_{M_i} = I_q$. Hence, for the full model we have no bias.

The estimated $MSE(\hat{\mu}_{n,M_i})$, $\widehat{MSE}_n(\hat{\mu}_{n,M_i})$, is found by replacing the parts of $MSE(\hat{\mu}_{n,M_i})$ with the corresponding estimates. The estimates for τ_0, ω, $G_{M_i}$and $Q^0_{M_i}$ are based on estimating J, which can be estimated without much problem with methods that are consistent and converges at normal rate, as described above in the discussion of the limiting distribution for MLE. For δ, however, it is no such estimator. Using $\hat{\delta}_{n,wide}$ is about the best we can do.[33] We see from equation (3.15) that what we need, is to estimate is $\delta\delta^t$. Since $E(DD^t) = \delta\delta^t + Q$, we use $\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}$, as an estimator for $\delta\delta^t$. The estimate becomes

$$\widehat{MSE}_n(\hat{\mu}_{M_i}) = \frac{1}{n}(\hat{\tau}^2_{0,n} + \hat{\omega}^t_n\hat{Q}^0_{M_i}\hat{\omega}_n + \hat{\omega}^t_n(I_q - \hat{G}_{n,M_i})(\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \hat{G}_{n,M_i})^t\hat{\omega}_n \tag{3.16}$$

$FIC(\hat{\mu}_{n,M_i i})$ is found by multiplying $\widehat{MSE}_n(\hat{\mu}_{n,M_i})$ by n

$$\begin{aligned} FIC(\hat{\mu}_{n,M_i}) &= n[\widehat{MSE}_n(\hat{\mu}_{n,M_i})] \\ &= \hat{\tau}^2_{0,n} + \hat{\omega}^t_n\hat{Q}^0_{n,M_i}\hat{\omega}_n + \hat{\omega}^t_n(I_q - \hat{G}_{n,M_i})(\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \hat{G}_{n,M_i})^t\hat{\omega}_n \end{aligned} \tag{3.17}$$

The estimated bias squared term $\hat{\omega}^t_n(I_q - \hat{G}_{n,M_i})(\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \hat{G}_{n,M_i})^t\hat{\omega}_n$ might be negative. There are several alternatives for avoiding such cases by a bias-modified FIC.[34] An alternative is to truncate this term, i.e., to replace this term by

$$max\{0, \hat{\omega}^t_n(I_q - \hat{G}_{n,M_i})(\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \hat{G}_{n,M_i})^t\hat{\omega}_n\}$$

As for AIC, FIC is more or less easily expanded to the regression setting. In our regression setting $\gamma_0 + \delta/\sqrt{n}$ will typically be the parameters associated with the covariates, i.e $\gamma_0 + \delta/\sqrt{n} = (\beta_0, \beta_1, ..., \beta_p)^T$, using the regression setting notation above. In the case of regression the density changes to

$$f_n(y) = f(y \mid x_i, \theta_0, \gamma_0 + \delta/\sqrt{n}) \tag{3.18}$$

By using the same principles as used in equation (3.5) we can construct a Fisher-information matrix J_n that converges to J.[35]

In the regression setting we can assume that we want to investigate if we should use model j, which add an additional covariate, which corresponds to adding a parameter in the GLM setting, to model i. The reduction in MSE by using model j is $\widehat{MSE}(\hat{\mu}_{n,M_i}) - \widehat{MSE}(\hat{\mu}_{n,M_j})$, and the corresponding estimate is

$$\widehat{MSE}_n(\hat{\mu}_{n,M_i}) - \widehat{MSE}_n(\hat{\mu}_{n,M_j}) = \frac{1}{n}(FIC(\hat{\mu}_{n,M_i}) - FIC(\hat{\mu}_{n,M_j}))$$

[33] See Claeskens and Hjort (2003) p. 150.
[34] See Claeskens and Hjort (2008) p 150.
[35] See Claeskens and Hjort (2008) p. 149.

Hence, if MSE is the expected loss of using a wrong model, then $\frac{1}{n}(FIC(\hat{\mu}_{M_i}) - FIC(\hat{\mu}_{M_j}))$ is an estimate of the value of information in the additional covariate in model M_j.

An observant reader might question the necessity of imposing the distributional assumptions used in the FIC framework. By using the delta theorem in and the properties of the MLE explained in section 3.2 above we get for a focus μ

$$\sqrt{n}(\mu_{M_i}(\hat{\theta}_{n,M_i}) - \mu(\theta_0)) \xrightarrow{d} N(0, (\nabla\mu)^t J^{-1} K J^{-1} (\nabla\mu))$$

where $\hat{\theta}_{n,M_i}$ are the MLE parameters under model M_i. By applying this formula we easily find the MSE of $(\mu_{n,M_i}(\hat{\theta}_{n,M_i}) - \mu(\theta_0))$ which is $\frac{(\nabla\mu)^t J^{-1} K J^{-1} (\nabla\mu)}{n}$, that can easily be estimated. But this would not help us much, because recall that θ_0 is the parameter that minimizes the K-L distance to the true DGP. Hence $\mu(\theta_0)$ is not the same as μ_{true} in the FIC framework. Thus, we will not get an estimate of the MSE with respect to the true focus as we get in the FIC framework.

3.3.3 Reduction in expected loss as a general framework for information value

It would be a coincidence if the expected K-L distance reduction or MSE reduction perfectly corresponded to the economic value of information from a better model in a specific situation. Hence, we cannot generally use changes in AIC or FIC as measurements for the the value of information. A more general approach is to assume that the value of information provided by a model can be measured as reduction in expected loss. The information value of a using a better model is the reduced expected loss associated with using the better model.

The expected loss approach to information value is justified by the rationality foundation of the expected utility paradigm developed by von Neumann and Morgenstern (1944) and later by Savage (1954) under more general circumstances allowing for subjective probabilities. They showed that under quite general conditions maximizing expected utility (or equivalently, minimizing expected loss) is equivalent to adhering to certain rationality axioms, such as transitivity of preferences and updating information according to Kolmogorovs rules of probability.[36]

Maximizing expected utility, and the equivalent, minimizing expected loss, is well established in statistical decision theory, and in decision theory more generally, as the foundation for rational decision making. Hence, there are qualified reasons to use expected loss of using a model as a criteria to be used in model selection, and the reduction of expected loss as the information value of using a better model. Both AIC and FIC as information criteria can be considered as special cases of expected loss minimization. In the application of AIC, the loss function is the K-L distance, and in the application of FIC loss function is the squared error. We will generally denote the expected loss of using model i as $EL(M_i)$. The information value, in terms of reduced expected loss, of using a model M_j instead of M_i is $EL(M_i) - EL(M_j)$.

[36] See Parmigani and Inoue (2009) for a comprehensive discussion.

In this study we rely on maximum likelihood estimators (MLE) in estimating the expected loss of a particular model. Hence, we will assume that the parameters of the models for which we want to estimate the expected loss, are estimated by MLE. This is a choice we have made, aware that it can be criticized on several grounds. One might ask if we are so concerned about minimizing some particular expected loss, then why not choose parameter estimates based on this particular expected loss function according to some criteria, for instance minimax (explained further below), and then measure the expected loss of a particular model using some criteria, for instance the estimated expected loss. We will not argue against other choices, but present some arguments in favor of our choice to use MLE.

The choice of using MLE can be defended by the nice statistical, and in particular, nice asymptotic properties of the MLE as described above. These nice properties make it easier to study properties such as variance and bias of an expected loss estimate, which is useful both for theoretical analysis and practical applications. There are also advantages to separate parameter estimation form the particular loss functions, because this makes it easier to use the same estimates under alternative loss functions. This can for instance be used to analyze the robustness of model selection to small variations in the loss function. Hence, there are good reasons to use MLE as parameter estimates in estimating the expected loss assocaited with a model.

Before we proceed with regression we will first assume IID, and then return to the regression setting. For the purposes of this study we will take the approach of Claeskens and Hjort (2003) assuming that we are interested in a focus μ which can be predicted or estimated under various models. $\hat{\mu}_{n,M_i}$ is the predicted/estimated μ based on the MLE when model M_i is used for estimation (and n is the number of observations). μ will be a function of model-parameters θ, i.e. $\mu(\theta)$. Since we don't know the parameters of the model, we have to estimate them from the data. $\hat{\mu}_{M_i}$ is based on the maximum likelihood parameters for M_i. Hence we can write $\hat{\mu}_{n,M_i} = \mu(\hat{\theta}_{n,M_i})$ where $\hat{\theta}_{n,M_i}$are the maximum likelihood parameters for M_i. This makes $\hat{\mu}_{n,M_i}$ the MLE under M_i due to the invariance property of the MLE. The loss will be the loss associated with the wrongful prediction or estimation of μ. We will then have

$$EL(M_i) = E[L(\mu, \hat{\mu}_{n,M_i})] \tag{3.19}$$

where $L(\mu, \hat{\mu}_{M_i})$ is the loss function associated with wrongful prediction/estimation. A special important case is $L(\mu, \hat{\mu}_{n,M_i}) = (\hat{\mu}_{n,M_i} - \mu)^2$, making the expected loss, $E[L(\mu, \hat{\mu}_{n,M_i})] = E[(\hat{\mu}_{n,M_i} - \mu)^2]$, the familiar MSE. We will discuss various loss functions in more detail in Chapter 5.

Since we don't know the true DGP to use in the calculation of the expected loss we must find an estimate for $E[L(\mu, \hat{\mu}_{n,M_i})]$. Let us assume that we want to estimate the expected loss according to the empirical distribution

$$E_n^*[L(\mu, \hat{\mu}_{M_i})] = \frac{1}{n}\sum_{j=1}^{n} L(\mu_j, \hat{\mu}_{n,M_i}) \tag{3.20}$$

where $L(\mu_j, \hat{\mu}_{n,M_i})$ is the observed loss associated with the j'th observation given that we operate with the model M_i.

$E_n^*[L(\mu,\hat{\mu}_{M_i})]$ is consistent, but we are interested in an unbiased estimate, $\widehat{E}_n[L(\mu,\hat{\mu}_{M_i})]$ since we are interested in correcting for sample bias. We need to correct for the fact that the same data are used to both estimate parameters and to estimate the losses. Adding an additional covariate will always reduce expected loss as we will get a better fit. Recall from above that the essence of deriving AIC was to adjust the empirical distribution, corresponding to $E_n^*[L(\mu,\hat{\mu}_{M_i})]$, for its bias. In the derivation of AIC it was possible to analytically derive a precise correction for bias. This is not an easy task for any loss function. Instead of trying to calculate a correction for bias, we might cope with the bias by using cross-validation (CV) in the estimation of $\widehat{E}_n[L(\mu,\hat{\mu}_{M_i})]$. Both the derivation of a bias correction for various loss functions, and the use of CV as an alternative will be crucial in Chapter 5 of this study.

Before we proceed, we will now discuss the inclusion of covariates in the expected loss estimation. Firstly, we must decide what the expected loss is in the context of covariates, i.e the covariates to be used in the expected loss function. Since we are going to predict some focus for a particular combination of covariates, let us say $x_0 = (x_{01},...,x_{0p})^t$, we should ideally get the expected loss for this particular combination of covariates. After all, we are ideally interested in $E[L(\mu,\hat{\mu}_{M_i}) \mid x_0)]$. However, this would violate the motivation for this study, as the question for this study is which covariates to gather. It would make non-sense to gather the covariates in the process of deciding what covariates to gather. Hence, we need to take the expectation over the variation of combinations of covariates that are likely to appear in a new subject.

To analytically derive the expected loss in this context one can assume that the covariates, as well as the responses are random variables, and take the expectation over the simultaneous distribution of responses and covariates.[37] Let us assume for a moment assume that that the covariates are random variables. We can then write equation (3.19) as

$$E[L(\mu,\hat{\mu}_{n,M_i})] = E_X[E[L_j(\mu,\hat{\mu}_{n,M_i} \mid X)]] \tag{3.21}$$

First we take the expectation of the expected loss with respect to $\hat{\mu}_{n,M_i}$ given certain values of the covariates, and then we take the expectation over the distribution of the covariates. This is, however, not the approach we will take here. Assuming the covariates to be random variables would complicate the analysis, since we will then also have to take into account sample biases with respect to the covariates in the estimation. Hence, in this study we will not consider the covariates as random variables.

Since we don't want to model randomness in the covariates, we have to make some assumptions. We simply narrow our expected loss to the expected loss associated with the covariates in the sample. Hence, we will instrumentally assume that we want to choose a model that minimize expected loss given for the covariate combinations in the sample. This means that we will be concerned in variations in the response, Y, for the covariates in the sample.

[37] See Claeskens and Hjort (2008) p. 25.

We will give each covariate combination equal weight 1/n. Hence, what we want to estimate is

$$E[L(\mu,\hat{\mu}_{M_i})] = \frac{1}{n}\sum_{j=1}^{n} E[L(\mu,\hat{\mu}_{n,M_i} \mid x_j)]$$

where $x_j = (x_{j0},...,x_{jp})^t$ are the covariates associated with observation j. By including including the covariates, equation (3.20) can be written as

$$E_n^*[L(\mu,\hat{\mu}_{M_i})] = \frac{1}{n}\sum_{j=1}^{n} L(\mu_j,\hat{\mu}_{n,M_i} \mid x_j) \tag{3.22}$$

We the get an estimate of average expected loss over various combinations of covariates. The challenge will be to find a $\widehat{E}_n[L(\mu,\hat{\mu}_{M_i})]$ that correct for possible bias in $E_n^*[L(\mu,\hat{\mu}_{M_i})]$. Since we want to estimate the expected loss over the covariates in the sample, this means we will have to correct only for in-sample bias. This means that we will only take into account the bias due to fitting the covariates in the sample to the observed responses, and not the bias resulting from having a particular selection of covariates among many potential combinations of covariates.

Estimating the expected loss associated with covariates observed in the sample seems to be a reasonable approach for the purposes of this study, which is model selection.[38] We just have to assume, in some sense, that the covariate combinations in the sample are "representative" for a future observation. How this actually will be implemented will depend on the context. In the derivation of AIC this approach is indirectly taken as it is the covariates from the sample that are used to estimate the log-likelihood function, giving each covariate combination equal weight. In the context of FIC we can solve this by averaging the FIC over all covariates, but a separate criterion is developed for this purpose, called AFIC (Average-FIC)[39] We will return to the precise derivation of AFIC below. Another justification for this assumption is that as n grows, any reasonable bias correction will converge to zero and $E_n^*[L(\mu,\hat{\mu}_{M_i})]$ will converge in probability to $E_X[E[L_j(\mu,\hat{\mu}_{n,M_i} \mid X)]]$ by the WLLN. Hence, in the limit, in-sample measurements also captures the distribution of covariates.

Using $E_n^*[L(\mu,\hat{\mu}_{M_i})]$ directly in the prediction or estimation of loss directly require direct observation or knowledge of the true parameters. Direct observation is available in prediction settings. Let say $\mu = Y_{new}$, i.e, our focus is to predict a new response. We then have

$$E_n^*[L(Y_{new},\hat{y}_{new,n,M_i})] = \frac{1}{n}\sum_{j=1}^{n} L(y_j,\hat{y}_{n,M_i,j} \mid x_j)$$

where $\hat{y}_{n,M_i,j}$ is the fitted y for a covariate combination x_j. This can be used as input in $\widehat{E}_n[L(\mu,\hat{\mu}_{M_i})]$, which also contain some bias correction term correcting for sample bias.

[38] Hastie et al. (2009) p. 230 argue for this approach to model selection in a regression context.

[39] See Claeskens and Hjort (2003) p. 179.

We will normally not have direct observation of a focus parameter to be estimated, such as the upper 5-precentile of a probability distribution, nor do we have the true parameters to calculate the true value. Hence, $E_n^*[L(\mu, \hat{\mu}_{M_i})]$ cannot be calculated directly. This can however be dealt with using the principles behind the derivation FIC, where the direct knowledge of the true parameters are avoided in estimating the expected loss. In the derivation of FIC the knowledge of the true parameters is avoided because we make an assumption regarding the nature of the true DGP which enable us to derive a limit distribution where the true parameter is included. A crucial part of this study will be to explore principles to extend the principles of FIC to estimate the expected loss associated with the candidate models under more general loss functions.

3.3.4 Risk function and Bayesian analysis

In the statistical inference literature, the expected loss function associated with estimation is often referred to as the risk function of the estimator. Hence, in the framework used above, we have that

$$R(\mu, \hat{\mu}_{M_i}) = E[L(\mu, \hat{\mu}_{M_i})]$$

where $R()$ is the risk function associated with $\hat{\mu}_{M_i}$ for a particular value of μ (and possible other variables). The risk function is useful for determining the risk of an estimator given the true parameter. This can be used to compare the risk of estimators over a range of possible parameters, often visualized in a figure. This is instructive to compare estimators. Typically it is found that some estimators performs better in one range of parameter-values, typically close to zero, while another estimator performs better in another range. The risk function is also useful for finding risk-motivated estimators such as the minimax-estimator, which minimizes maximal risk.

We have chosen to mainly stick to the expected loss terminology instead of using the term risk. The main reason for this is that risk is often, as just mentioned, used to compare various estimators, for instance MLE with some other estimator. This is not the approach taken in study, where we want to estimate the expected loss of using a particular model for a data-set where the parameter is estimated by MLE. An additional reason for this is that the term risk has several meanings in different parts of statistical literature, while expected loss is less ambiguous. Furthermore, we think it is more intuitively appealing to think of the value of information of a model in terms of reduced expected loss rather than reduced risk.

The risk function is particularly useful in Bayesian analysis and decision theory, where a probability distribution is associated with an unknown parameter θ. If a parameter θ is estimated by $\hat{\theta}$, we have that $E_{\pi(\theta)}[R(\theta, \hat{\theta})]$ is the expected risk of $\hat{\theta}$, when parameter probabilities are taken into account. This risk is known as Bayes risk. A rational parameter estimation choice would be to minimize this risk. The estimator that minimizes Bayes risk is known as the Bayes estimator. This study is frequentistic in the sense that we will not associate probability distributions to parameters.

3.4 Cost of gathering covariates

In the regression and model selection literature, the costs associated with gathering covariates normally do not play a big role. If we for instance are doing parameter inference or want to find AIC for various models for a given set of data, the covariates are already there to the statisticians disposal. However, if we want to do predictions for a focus, for instance to predict Y for a new set of covariates, we must gather the new covariates to use for prediction.

There can be substantial costs associated with gathering covariates. There are several practical examples of such a setting. In medical statistics, we might have several measurements (covariates) for historical patients. However, a measurement associated with a new patient is associated with costs. In financial analysis and risk management, we might have historical publicly available accounting information for some firms available at a low costs. However, to obtain the most recent information, high cost may have to be incurred. In a sea rescue operation, meteorological analysis is crucial, but getting the data, and the time needed to get the data, are associated with high costs. It is not hard to come up with examples where there are costs associated with gathering data. Actually, it is probably rather the rule than the exception that there are costs associated with gathering data.

Constructing cost functions is within the domain of economics.[40] Economic theory provides insight into the main characteristic of cost functions. A discussion on the construction of possible covariate cost functions and how covariate costs are related will be provided in Chapter 4. However, the construction of cost functions will not play a main role in this study. We will mostly assume the cost associated with gathering the relevant covariates for model i, $C(M_i)$, to be given. In some cases it will be possible to calculate the difference in costs associated with two alternative models i and j, $C(M_i) - C(M_j)$, even if the total costs of using a model cannot be easily calculated. For instance, if one covariate is what distinguishes two models, we only need the cost of gathering that particular covariate to make a cost-information trade-off. This will be elaborated further in the next subsection on optimization.

Costs may be subject to random influences. Statistics and econometrics can be used to calibrate the cost models with data and perform inferences. We will generally assume cost to be deterministic, as our focus will be on the estimation of the information-value of covariates to be traded off against costs. However, we will briefly discuss the extensions necessary to take into account random influences on costs in Chapter 4.

Note that a similar, but still different issue, issue is the costs of gathering data in the first place. By gathering more data one normally get more precise estimates, which has a value. If an experiment is performed it is an economic issue how many test subject to include in the experiment. This is in particular known to those who are involved in the arrangement of case-control studies, cohort studies and choosing between such studies where both are applicable. In a case-control study it is often substantial costs associated with the investigation of each person to include in the study for retrospective analysis. In a cohort study there are often substantial costs associated with each individual to follow prospectively,

[40] See for instance Varian (1992) chapters 4 and 5.

especially if each of the individual is to be exposed to some treatment or medication. In a market survey it is an economic question how many subjects to interview. With more data, we normally can make more precise inferences and get results of higher power. The trade-off in how much data to collect to make inferences is a complementary problem to the problem addressed in this study. In this study the data are assumed given, and the trade-off is which covariates that should be gathered for the prediction regarding a new subject. An interesting extension of this study is to take both trade-offs into account.

3.5 Optimization

Equipped with tools for how to determine the value of information from adding covariates to a model in the prediction or estimation of a focus, and knowing the costs associated with gathering covariates, we can trade-off the information-value and costs. The total cost associated with a model is the expected loss associated with the model plus the cost of gathering covariates. Hence the total costs of applying model M_i is

$$E[L(\mu, \hat{\mu}_{M_i})] + C(M_i)$$

Both components should be taken into account in rational model selection. The model that minimizes the total cost should be chosen.

Since we don't know the exact expected loss of using a model as the true DGP is unknown, we have to operate with an estimate. Hence the total estimated costs of applying M_i, is

$$\widehat{E}_n[L(\mu, \hat{\mu}_{M_i})] + C(M_i)$$

In the question on whether to include an additional covariate or not we just need to know the reduction in total costs from including this covariate. Assume that model M_j adds a covariate to model M_i. The total cost reduction is given by

$$\Delta = (E[L(\mu, \hat{\mu}_{M_i})] - E[L(\mu, \hat{\mu}_{M_j})]) - (C(M_j) - C(M_i))$$

We see that $\Delta > 0$ if and only if $E[L(\mu, \hat{\mu}_{M_i})] - E[L(\mu, \hat{\mu}_{M_j})] > (C(M_j) - C(M_i))$. Hence the information-value of including the covariate must exceed the incremental cost of including the covariate. Since we only have estimates for the expected loss, we must investigate whether

$$\widehat{E}_n[L(\mu, \hat{\mu}_{M_i})] - \widehat{E}_n[L(\mu, \hat{\mu}_{M_j})] > C(M_j) - C(M_i)$$

The trade-off between information-value and costs will be illustrated further by a simulation experiment in Chapter 6.

Remark 3.1. In Chapter 2 we referred to Raiffa and Schlaifer (1961), who in the Bayesian setting (without model uncertainty) proved that you should never say no to free information. Simplified one can say that if you are in a setting with a loss function $L(X)$, where X is some random event with distribution function

$f(x \mid \theta)$ and a prior $\pi(\theta)$, you will, a priori, face a lower expected loss, if you gather the value of θ rather than relying on the prior alone. In our setting, where the question is to gather a covariate for estimation or prediction, it might not reduce expected loss to take this covariate into account. The reason is that this involves using a model incorporating this covariate, which might be subject to a higher expected loss. This is most easily seen in the mean squared error loss framework. Including an additional covariate in a model is likely to reduce the bias squared of the prediction or estimate, but the cost is a higher variance. The impact of the variance is higher, the lower the sample size. Hence, it might not be rational to gather a covariate even if it is costless. This will be illustrated in our simulation experiment in Chapter 6. Strictly speaking, however, the statement of Raiffa and Schlaifer (1961) still holds, because even if we have a covariate, we are free to not use it. Hence, we cannot do worse by gathering it.

4 Covariate cost functions

In this chapter we will elaborate on how covariate cost functions might look like. As mentioned in the analytical framework in Chapter 3, the construction of cost functions is not a main issue for this study. However, this chapter is meant to provide the reader with some guidance on how to construct cost functions associated with gathering covariates. We will first present a covariate cost framework where the costs of gathering covariates are linear and independent of each other. After this, we will present some extensions to this framework, taking into account that the costs of gathering the different covariates can be interrelated. Finally, although we will assume deterministic cost functions in study, we will discuss reasons why the costs might be associated with random influences and the need to estimate cost functions.

4.1 Simple linear covariate costs

Recall from above that a regression model can be written as $Y \sim f(y \mid x_0, ..., x_p; \theta)$. Furthermore, recall that $M_{\{klm\}}$ corresponds to the the model where the covariates k,l and m are included. For instance $M_{\{012\}}$ corresponds to the model $Y \sim f(y \mid x_0, x_1, x_2; \theta)$. The cost of gathering the covariates k,l,m assuming linear independent costs, can be written as

$$C(M_{\{klm\}}) = C(x_k, x_l, x_m) = C(x_k) + C(x_l) + C(x_m) = c^t 1(M_{\{klm\}}) \tag{4.1}$$

$c = (c_1, ..., c_p)^t$ is a vector of constants and $1(M_{\{klm\}})$ is an indicator vector with one-components corresponding to the covariates that are included in the model. In this cost function it is assumed that there is a fixed constant cost associated with gathering a covariate independent on whether other covariates are gathered.

It is not difficult to imagine that the cost assumptions above might be violated, but still assuming independence. The cost of gathering a covariate could depend on the value of the covariates. It is for instance possible to think that it is more costly to measure the speed of some particle the higher the speed

(or opposite). Including such a possibility would cause many complications. If the cost is dependent on the value of the covariate, a question is how this should be dealt with ex ante. One possibility is that one first costless could determine which interval the value is within. However, then one already has valuable information that may not make it worth it to gather the exact value of the covariate. Another possibility is that there could be a probability distribution associated with the value of the covariate that could be taken into account in the cost assessment. Then one get an additional uncertainty that must be included in the analysis.

One could also assume an extension where the costs increases the more precise you want to measure the value of the covariate. This would probably apply to many real world situations. It is, for instance, probably more costly to measure the speed of a vehicle, the more precise you need the measurement to be. We will explore this issue further in the simulation experiment in Chapter 6, where we will assume that a covariate is a function of another covariate plus noise, but that the noisy version is cheaper to observe.[41]

4.2 Economies of scope and sequence

Above it was assumed that the costs of gathering covariates were independent of each other. This might however be in conflict with many real world situations. Often, there will be some economies of scope. If, for instance, a person is interviewed, it will be a small additional cost of asking the person about an additional question, for instance, whether he smokes or not. The main cost is the fixed cost of establishing the interview. An obvious example of economies of scope in the statistical setting is the economies of scope in gathering polynomials of the same covariate. If the covariate x_i is already gathered, it is costless to "gather" the polynomial x_i^m for any i. Economies of scope can be described with the following notation. We can say that there are economies of scope in gathering covariate k and l if

$$C(M_{\{kl\}}) < C(M_{\{k\}}) + C(M_{\{l\}}) \tag{4.2}$$

Hence, economies of scope in gathering covariate k an l means that it it are less costs associated with gathering both covariate k and covariate l together, than the sum of gathering the costs individually. Another way to see equation (4.2) is to rearrange to

$$C(M_{\{kl\}}) - C(M_{\{k\}}) < C(M_{\{l\}})$$

This means that the additional cost of gathering l given that one also gather k is less than the individual cost of gathering covariate l.

Note that in the simple example above there are no dynamical considerations. One assume that both are gathered. Hence, there is symmetry in the meaning that cost reductions applies independent of which of the covariate gathered first. We also have $C(M_{\{kl\}}) - C(M_{\{l\}}) < C(M_{\{k\}})$, which means that

[41] This has parallels to state-space models used in the the analysis of time-series.

the additional cost of gathering k given that one also gather l is less than the individual cost of gathering covariate k.

The symmetry in economies of scope might be in conflict with many real world applications. If we introduce dynamic considerations one could for instance imagine that there is some economies of scope in gathering covariate l given that k is gathered, but not opposite. In other words one can think that the cost of gathering k is independent on whether l is gathered or not, but given that k is gathered it is less costly to gather l. This can be described as $C(M^{(kl)}_{\{kl\}}) < C(M_{\{k\}}) + C(M_{\{l\}})$, but $C(M^{(lk)}_{\{kl\}}) = C(M_{\{k\}}) + C(M_{\{l\}})$. The superscript (kl) in $M^{(kl)}_{\{kl\}}$ means that the covariates are gathered in the sequence k-l. An application of this can be that it is performed an expensive and extensive test on a subject to gather covariate k. Given that this test is performed it is relatively low cost associated with extracting the value of the covariate l. However, if we only were to gather covariate l, we would perform a simpler less costly test, but still more costly than if the extensive test was already performed.

4.3 Some words on costs subject to random influences

Above it was assumed that the costs were deterministic. In real world applications there are stochastic influences that makes the costs random. Fluctuating prices on inputs, such as electricity prices, can affect the cost of gathering covariates. Some costs of gathering covariates are likely to be weather-dependent and so on. Costs subject to random influences would mean that we would have to take the uncertainty properly into account from a decision theoretic perspective. The information-value of covariates measured by the loss due to wrongful estimation cannot just be traded of against the best cost estimates. Rather we should incorporate the cost into our total expected loss function. Due to risk aversion we might, for instance, discard a good model with relatively low estimated costs, if there is a small, but positive probability of particular high costs associated with gathering a covariate included in the model. For instance, let us say that we need some helicopter to perform a measurement to be used in a weather forecast model. The expected cost of performing such measurement might not be extremely high, but a low risk of accidents may prevent the rationality of gathering this covariate.

When the costs are subject to random influences estimation might be necessary as the parameters and even the structure of the cost model itself, might not be known. Econometric analysis can be used to estimate parameters in cost model subject to random influences. Cost estimates, such as $\widehat{C}(M_{\{klm\}})$ can be obtained.

In this study we want to pay our attention to measuring the information-value part of the covariates, and covariate costs are more considered as a reason for the necessity of measuring the information-value of covariates. Hence, we will not elaborate into cost estimation and associated issues. However, a natural extension to this thesis would be to allow for random influences on the costs.

5 Loss functions and estimating the information-value of covariates

In this chapter we will go into the details of alternative ways to estimate expected loss to approximate the information-value of regression covariates, i.e the information-value of a better model. We will start by a discussion on how to construct reasonable loss functions for the purpose of measuring the information-value of covariates in terms of expected loss reduction. This will be followed by a short discussion on the relationship with covariates, parameter complexity and information-value in regression settings.

Next we will discuss how to directly estimate the information value of a model and its associated covariates. The main challenge in this part will be to construct unbiased estimates of the information value based on MLE plug-in and the empirical distribution function. Normally we will have to add a bias correction term on the MLE plug-in empirical estimators. The bias-correction must be done because of sample bias, i.e. we use the same sample to estimate estimators and to estimate the expected loss associated with a model.

A way to avoid or reduce sample bias is to use cross-validation (CV) in the expected loss estimation, which will be discussed next. By employing CV, we separate the data used for parameter estimation and expected loss estimation. A disadvantage of the CV approach is that it often requires extensive programming and often time consuming computation resources.

Both direct estimation and CV are limited to situations where the loss function is a function of observable variables. In other words, the focus must be a function of observables. For instance, if our focus is to predict a new response, we have the responses in the sample to estimate the loss. In parameter estimation settings we don't any observations of the true focus. In this setting, i.e. the setting where the focus is a parameter, we will explore how far FIC can help us in estimating the expected loss associated with other loss functions.

5.1 Constructing loss functions for estimating economic information-value

5.1.1 Economic information value versus statistical information-value

Many loss functions can be defended on statistical grounds. The K-L distance as an expected loss function has information-theoretic justifications and is crucial to understanding the statistical appealing properties of the MLE. The mean squared error has geometrical justifications, and allow us, for instance, to see regression as projections.

However, in this study we are not mainly concerned with the statistical properties associated with using a particular loss function. Since our ultimate goal is to trade-off information-value with costs, we are interested in the economic information value of a regression model and its covariates. The economic information-value of a model depends on the expected loss in the context the model is chosen within. For investors, a better model allow for better economic estimations, which obviously have economic value. In medical diagnostic settings accurate estimation can in the extreme case be a question of life an death. In this case the economic value of better estimation in this case must be indirectly constructed by economic

methods. In weather forecasts, it is easy to see that better predictions may reduce costs, for instance in capacity planning.

Below we aim to give some direction on how to construct an appropriate loss function when the purpose is to estimate the economic value of information provided by a model. Such loss functions must necessarily be on a monetary scale. This means that even if traditional statistical loss function such as K-L distance and squared error are appropriate and are used, they must necessarily be calibrated to be on a monetary scale.

5.1.2 The general properties of reasonable loss functions

In the decision theory and economics literature it is normally separated between risk aversion, risk neutrality and risk loving. It is usually assumed with empirical support that individual persons are risk averse. This means that the increased utility of more wealth is decreasing. In mathematical terms, assuming smooth utility functions, this means that $U'(x) > 0$, but $U''(x) < 0$, where U is the utility function and x is some positive wealth. Transferred to a loss setting this means that $L'(y) > 0$ and $L''(y) > 0$, where y now is some positive loss. This corresponds with many loss functions traditionally used in statistics such as the squared error loss function, where $L'(y) = 2y$ and $L''(y) = 2$. Companies, on the contrary to individuals, should be risk neutral as investors more effectively can satisfy their (possible risk averse) risk preferences by designing a suitable portfolio of stocks of various risk rather than through the investment in a single firm.[42] This means that $L''(y)$ is zero. It seems reasonable to assume that when we construct real world loss functions for economic evaluations, they should either reflect risk aversion or risk neutrality depending on the circumstances.

In this study we are interested in economic losses associated with statistical estimation. The loss functions are associated with the errors in estimation and prediction. It is not possible to give a general answer to how a loss function should look like in this setting. In some cases risk aversion seems to be reasonable. Risk aversion probably applies in several environmental settings. A wrongful estimation of the cleaning capacity of an environment can have catastrophic consequences. Hence, risk averseness should apply in the choice of model for the estimation of cleaning capacity. In financial settings risk neutrality is probably often the case. In estimating the expected return of a investment project within a company it can probably be argued for risk neutrality. Hence, risk neutrality should apply in selecting the model for estimating the return. For other focuses, however, risk aversion is probably more representative for the actual loss function. In financial risk management, Value at Risk (VaR) is often used as a risk measure in risk reporting and risk management.[43] If Y_T is the profit a over a time horizon T, VaR_α is the

[42] This is a general result from financial analysis, see for example Posner (2011) § 15.1.

[43] The use of VaR as a risk measure to compare risks can be criticized on several grounds, inter alia, for violating rationality criteria. See Beneplanc and Rochet (2011) p. 67. It will be beyond the scope of this study to elaborate into this discussion.

value such that the probability that Y_T is less than VaR_α is α.[44] In mathematical terms, this means that

$$P(Y_T \leq VaR_\alpha) = \alpha$$

which means that $VaR_\alpha = F^{-1}(\alpha)$. Since VaR is often negative it is usually described in term of L_T, such that $L = -Y_T$. The VaR could for instance be that there is a 5 percent chance that the loss from a financial portfolio is greater than 100 million dollar. Hence minus 100 million dollar is the 5 percent percentile of the profit distribution. VaR is typically used as a proxy to assess risk, and it is arguable that risk aversion should apply in choosing a model for the estimation of VaR.

A loss function that is quite general of nature, but still quite easy to work with, is one where the the loss of wrongful estimation of a focus is dependent on the difference between the true focus and the estimated focus (error). We can write this loss function as

$$L(\mu, \hat{\mu}_{n,M_i}) = g(\hat{\mu}_{n,M_i} - \mu)$$

μ is the true value of the focus and $\hat{\mu}_{n,M_i}$ is the estimated focus using model M_i. The function g is general and allow for possible asymmetric loss functions in the meaning that overestimating μ might be worse than underestimating μ and vice versa. A familiar function with symmetric loss is $g(x) = x^2$ which is the squared loss. This loss function satisfies a requirement of risk aversion. The loss $g(x) = |x|$ is also symmetrical and reflects risk neutral preferences. We will return to these and other loss functions below.

The reduced loss from using a model M_j instead of M_i is

$$L(\mu, \hat{\mu}_{n,M_i}) - L(\mu, \hat{\mu}_{n,M_j}) = g(\hat{\mu}_{n,M_i} - \mu) - g(\hat{\mu}_{n,M_j} - \mu)$$

5.1.3 Some examples of particular loss functions and economic loss calibration

There are several loss functions used in the literature. As mentioned above a loss function widely used in statistics is the squared error,

$$L(\mu, \hat{\mu}_{n,M_i}) = (\hat{\mu}_{n,M_i} - \mu)^2 \tag{5.1}$$

An obvious way to calibrate this loss with some real world economic loss would be to multiply the squared loss function with some constant b, hence $L(\mu, \hat{\mu}_{n,M_i}) = b(\hat{\mu}_{n,M_i} - \mu)^2$. An extension to this is to add a constant a, with resulting loss function $L(\mu, \hat{\mu}_{n,M_i}) = a + b(\hat{\mu}_{n,M_i} - \mu)^2$. This loss function complies with risk averse preferences. More generally, one imagine other transformations using the squared loss function as a basis. Hence, we could create loss functions of the type

$$L(\mu, \hat{\mu}_{n,M_i}) = g((\hat{\mu}_{n,M_i} - \mu)^2)$$

An example of this last type of loss function could be $g((\hat{\mu}_{n,M_i} - \mu)^2) = ae^{b(\hat{\mu}_{n,M_i} - \mu)^2}$. The risk preferences implied by $g()$, depends on the particular formulation of $g()$. $g()$ could be constructed to take into account assymetric preferences in the direction for the error. For instance, we could let

[44] Loosely based on Beneplanc and Rochet (2011) p. 55.

$$L(\mu,\hat{\mu}_{n,M_i}) = \begin{cases} a+b_1(\hat{\mu}_{n,M_i}-\mu)^2 & \text{if } \hat{\mu}_{n,M_i} \geq \mu \\ a+b_2(\hat{\mu}_{n,M_i}-\mu)^2 & \text{if } \hat{\mu}_{n,M_i} < \mu \end{cases}$$

Another loss-function which is familiar to most statisticians is the absolute error loss function

$$L(\mu,\hat{\mu}_{n,M_i}) = |\hat{\mu}_{n,M_i}-\mu| \tag{5.2}$$

As for the the squared loss function this loss function can be calibrated to some real world economic loss by letting $L(\mu,\hat{\mu}_{n,M_i}) = a+b|\hat{\mu}_{n,M_i}-\mu|$. This loss function reflects risk neutral preferences. More generally we can let $L(\mu,\hat{\mu}_{n,M_i}) = g(|\hat{\mu}_{n,M_i}-\mu|)$. $g()$ can as for the squared error be constructed to allow for asymmetric preferences in the direction of the error.

Another loss function widely used by statisticians is the zero-one loss given by $L(\mu,\hat{\mu}_{n,M_i}) = I(\hat{\mu}_{n,M_i} \neq \mu)$. This is also know as the indicator loss function. The loss is incurred only if μ is wrongly predicted and, if the loss is wrongly predicted, it does not matter how much it is wrongly predicted. Hence, the indicator reflects risk neutral preferences. As for the squared error and absolute error, the indicator loss can be calibrated to reflect some real economic loss. An obvious candidate is $L(\mu,\hat{\mu}_{n,M_i}) = a+bI(\hat{\mu}_{n,M_i} \neq \mu)$, or more generally $L(\mu,\hat{\mu}_{n,M_i}) = g(I(\hat{\mu}_{n,M_i} \neq \mu))$. An apparent application of the zero-one loss function is in classification problems. If you predict correct class there is no loss, else there is a loss. In case of classification the loss can be made asymmetric dependent on which class that is wrongly predicted. Assume for instance that a binomial regression is used to predict classes 0 and 1. we could then write

$$L(\mu,\hat{\mu}_{n,M_i}) = L_1 I(\hat{y}=0 \mid Y=1)) + L_0 I(\hat{y}=1 \mid Y=0)$$

where $\hat{y}$ is the predicted class, L_1 is the loss of misclassifying class 1 as class 0 and L_0 is the loss of misclassifying class 0 as class 1. Another apparent situation where the indicator loss applies is in betting situations. Either you estimate exactly, or you loose. This would for instance apply in the prediction of a financial asset price for the purpose of of deciding whether to buy a a call option.

A loss function closely related to the zero-one loss function, but less restrictive, is the loss function where the loss is zero if the wrong prediction is within certain limits. More precisely, we can let $L(\mu,\hat{\mu}_{n,M_i}) = I(|\hat{\mu}_{n,M_i}-\mu| > \varepsilon)$. An observation is that the expected loss in this situation turns into a probability as $E[I(|\hat{\mu}_{n,M_i}-\mu| > \varepsilon)] = p(|\hat{\mu}_{n,M_i}-\mu| > \varepsilon)$. More generally, we could let $L(\mu,\hat{\mu}_{n,M_i}) = g(I(|\hat{\mu}_{n,M_i}-\mu| > \varepsilon))$.

Another loss function used in statistics is the LINEX (Linear-Exponential) loss function developed by Varian (1974) and further by Zellner (1986). The LINEX loss function is

$$L(\mu,\hat{\mu}_{n,M_i}) = g(\hat{\mu}_{n,M_i}-\mu) = b(e^{a(\hat{\mu}_{n,M_i}-\mu)} - a(\hat{\mu}_{n,M_i}-\mu) - 1) \tag{5.3}$$

For the LINEX loss function to have meaning we must have $a \neq 0$. For losses to be positive we must have b>0. We see that b works as a scale parameter, that can be used to calibrate the loss function to a suitable level for the economic loss for the problem we are analyzing.

Note that as $a \to 0$, LINEX reduces to a calibrated squared error loss function. This is most easily seen by a Taylor development of e^z around 0 which gives us

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + ... = \sum_{k=1}^{\infty} \frac{z^k}{k!}$$

Hence $e^z - 1 - z = \frac{z^2}{2!} + \frac{z^3}{3!} + ... = \frac{z^2}{2!} + R$. As $z \to 0$, the first term will dominate the remainder, R. Hence, as $a \to 0$, we have

$$\begin{aligned} L(\mu, \hat{\mu}_{n,M_i}) &= b(e^{a(\hat{\mu}_{n,M_i} - \mu)} - a(\hat{\mu}_{n,M_i} - \mu) - 1) \\ &= b(1 + a(\hat{\mu}_{n,M_i} - \mu) + \frac{a^2(\hat{\mu}_{n,M_i} - \mu)^2}{2!} + \frac{a^3(\hat{\mu}_{n,M_i} - \mu)^3}{3!} + ... - a(\hat{\mu}_{n,M_i} - \mu) - 1) \\ &= b(\frac{a^2(\hat{\mu}_{n,M_i} - \mu)^2}{2!} + \frac{a^3(\hat{\mu}_{n,M_i} - \mu)^3}{3!} + ...) \end{aligned}$$

which converges to $b\frac{a^2(\hat{\mu}_{n,M_i} - \mu)^2}{2}$ as $a \to 0$. A remark in this context is that if we re-parametrize and replace b by $b_0 = \frac{2b}{a^2}$, then the LINEX loss function will converge to exactly the squared error loss as $a \to 0$. If this is done we get a useful recalibration where $a \to 0$ all the time converges to the squared error. Hence, we will in some sence decouple a and b. This recalibration might, for instance, be useful in sensitivity analysis.

The LINEX loss function has several nice properties. Firstly, it allows for asymmetric losses. As it is not trivial how the LINEX loss function looks like, it is illustrated in Figure 5.1. As we see, the LINEX loss punishes positive values more than negative values for $a > 0$, while opposite for $a < 0$. We can adjust the relative loss of positive values to negative values by adjusting a. We could also add a constant c to calibrate the loss further to the economic loss in a particular situation, i.e.

$$L(\mu, \hat{\mu}_{n,M_i}) = c + b(e^{a(\hat{\mu}_{n,M_i} - \mu)} - a(\hat{\mu}_{n,M_i} - \mu) - 1) \tag{5.4}$$

Another nice property of the LINEX loss function is the smoothness, which make it suitable for Taylor-developments. This will be returned to below.

Another loss function used in statistics is the Stein's loss function.[45] Transferred to our setting this loss function would be

$$L(\mu, \hat{\mu}_{n,M_i}) = \frac{\hat{\mu}_{n,M_i}}{\mu} - 1 - \log\frac{\hat{\mu}_{n,M_i}}{\mu}$$

This loss function has been pointed out as particularly useful in estimating parameters that must be positive (such as variance), because it penalizes gross over-estimation and gross under-estimation equally

[45] See Casella and Berger (2001) p. 351.

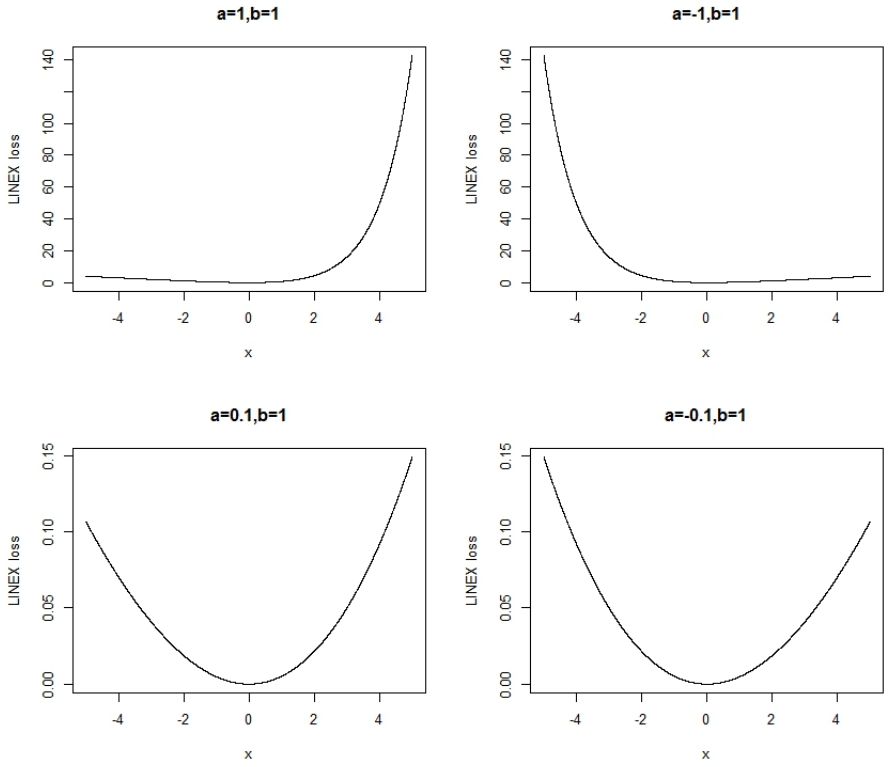

Figure 5.1: The LINEX loss

hard.[46] Other loss functions, such as the squared error, penalizes over-estimation more than under-estimation, since the over-estimation may potentially be infinite while under-estimation is bounded below by zero in case of parameters that can only be positive. We will not pursue this loss function in more detail here, but note that this is a loss function with desirable properties that probably, with some additional calibration parameters, also could be useful as a loss function for the economic loss in practical settings.

5.1.4 Finding the loss-function parameters

Specifying the parameters of the loss function is not always obvious. In some situations, such as financial settings, the loss is monetary as such, and the parameters follows naturally. In other situations it is easy to transfer the loss to some monetary scale because the losses are very close to being economic. If we, for instance, need a weather prediction model for planning snow cleaning, we normally have some idea of the cost on calling in extra personnel on short notice in case of wrong prediction, and/or costs related to delays and accidents due to lack snow cleaning. In some cases economics provide guidance by the principle of revealed preferences. If we, for instance, wrongly predict the self-cleaning capacity of the river, causing all the fish to die, the expenses incurred by sport-fishers to use the river provide some guidance on the value of the fish in the river. Deriving utility functions, and equivavlently, loss functions, is within the domain of economics and can be hard to do in practice.[47]

A point to be made is that finding the loss functions applicable to the relevant situation is something the statistician will have to do together with an expert in the relevant field, and possibly with the aid of an economist.The statistician and other experts have to sit down and figure out what kind of loss function is a reasonable approximation to the relevant situation, and the corresponding parameters. The statistician will typically provide guidance on what loss functions are available, and advantages and disadvantages associated with various loss functions from a statistical point of view. The statistician will also typically have a role in performing a sensitivity analysis for the choice of parameters.

Another point to be made is that in some circumstances the parameters of the loss functions may need to be estimated by statistical methods. In that case another layer of uncertainty is imposed on the analysis that ideally should be taken into account when taking rational decisions. We will not go into the details of this issue here, but simply assume that the loss function parameters are given.

Note also that from a statistical point of view it can be interesting and instructive to perform various analyses to gain insight on the choice of parameters. For instance, as mentioned above as $a \to 0$, LINEX reduces to a calibrated mean squared error loss function. An instructive analysis would be to analyze if there are situations where it is better to operate with MSE instead of LINEX for sufficiently small a, because of the analytical advantages associated with MSE. In practical situations a typical role of the statistican would be to assess, in the specific case, whether a is small enough to justify the use of the simpler MSE instead of the LINEX loss function.

[46] See Casella and Berger (2001) p. 351

[47] See for instance Varian (1992) Chapter 7.

5.2 Covariates, parameter complexity and information value

So far we have been rather quiet on estimating the information value of a regression covariate. In Chapter 3 we implicitly assumed that adding a covariate was equivalent to adding one parameter. As we will see this is the case for the GLM models, but not necessarily the case in more complicated settings.

As we saw GLM are characterized by the following distribution :

$$Y_i \sim f(y_i; \theta_i, \varphi)$$

where $f(y_i; \theta_i, \varphi)$ belongs to the overdispersed exponential family, with $\mu_i = E(Y_i) = b'(\theta_i)$ for a function $b(\theta_i)$. See Section 3.1.1 for details.

The covariates enters the distribution with $\eta = \beta + \beta_1 x_1 + ... + \beta_p x_p$ in the way that the mean μ_i is a smooth i invertible function of η. That means that we can write

$$\mu_i = m(\eta_i)$$

The link function is given by

$$\eta_i = m^{-1}(\mu_i) = g(\mu_i)$$

As we can see adding another covariate x_i in the GLM setting implicitly means adding another parameter β_i. Hence, when punishing the increased complexity in model selection, adding a covariate to the regression model can be considered as adding one parameter.

If we look beyond the GLM setting it is easy to construct models where adding a covariate might add two parameters to the model. Assume for instance we use a model of the type

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i$$

where residuals ε_i, i=1,...,n are independent distributed following a normal distribution $N(0, \sigma_i^2)$ with $\sigma_i = e^{\gamma_0 + \gamma_1 x_{i1} + ... + \gamma_p x_{ip}}$. We see that including x_i in this model involves adding two parameters: β_i and γ_i.

As stressed earlier, we will assume that we are within a GLM setting in this study, and hence assume that adding a covariate is equivalent to adding a parameter. However, the analysis can easily be expanded to the situation where adding a covariate might add more than one parameter.

5.3 Direct estimation of expected loss in prediction settings

5.3.1 General approach

In this section we will see how we can construct an unbiased estimate of $E[L(\mu, \hat{\mu}_{M_i})]$ by using the empirical distribution function. In this approach we are dependent of observations. Hence, this is a suitable approach in prediction settings where we can utilize the observations in the sample and the fitted values to obtain an estimate of the expected loss.

The empirical distribution direct plug-in estimate will be consistent according to the WLLN, but might be biased because of sample bias. The challenge will be to correct for bias. Let

$$E_n^*[L(\mu,\hat{\mu}_{n,M_i})] = \frac{1}{n}\sum_{j=1}^{n} L(\mu_j,\hat{\mu}_{n,M_i} \mid x_j)$$

be direct plug-in empirical estimate of the expected loss based on the empirical distribution function. Experience tells us that the unbiased estimate of $E[L(\mu,\hat{\mu}_{M_i})]$, at least approximately, can be written as

$$\begin{aligned}\widehat{E}_n[L(\mu,\hat{\mu}_{n,M_i})] &= E_n^*[L(\mu,\hat{\mu}_{n,M_i})] + z(|M_i|) \\ &= \frac{1}{n}\sum_{j=1}^{n} L(\mu_j,\hat{\mu}_{n,M_i} \mid x_j) + z(|M_i|)\end{aligned}$$

where $|M_i|$ is the dimension, i.e., the number of parameters in model i. $z(|M_i|)$ is the complexity punishment, so $z(|M_i|)$ is a positive, non-decreasing function. The more complexity, the higher is the bias from sample effects.

We cannot generally and easily calculate what $z(|M_i|)$ should be in all cases. It will depend on the specific loss function. However, we have some particular cases where it is fairly straight forward to calculate $z(|M_i|)$. Below we will go through some of these particular cases.

5.3.2 Squared prediction error as loss function

Assume now that

$$E[L(\mu,\hat{\mu}_{n,M_i})] = E[(Y_{new} - \hat{y}_{n,M_i,new})^2]$$

where Y_{new} is some unknown new realization of the response variable and $\hat{y}_{n,M_i,new}$ is the prediction of Y_{new} using model M_i. Furthermore, let $\widehat{f}_{n,M_i}(x_j)$ be the prediction of Y_{new}, given covariate combination x_j.

We want an unbiased estimate of

$$E[(Y_{new} - \hat{y}_{n,M_i,new})^2] = \frac{1}{n}\sum_{j=1}^{n} E[(Y_{new,j} - \widehat{f}_{n,M_i}(x_j))^2]$$

where $Y_{new,j}$ is a new realization, given covariate combination x_j. To find the bias of the the plug-in empirical estimate

$$E_n^*[(Y_{new} - \hat{y}_{n,M_i,new})^2] = \frac{1}{n}\sum_{j=1}^{n} (y_j - \widehat{f}_{n,M_i}(x_j))^2$$

we use that

$$
\begin{aligned}
& E[E_n^*[(Y_{new} - \hat{y}_{n,M_i,new})^2]] - E[(Y_{new} - \hat{y}_{n,M_i,new})^2] \\
= \; & E[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \widehat{f}_{n,M_i}(x_j))^2] - \frac{1}{n}\sum_{j=1}^{n}E[(Y_{new,j} - \widehat{f}_{n,M_i}(x_j))^2] \\
= \; & \frac{1}{n}\sum_{j=1}^{n}\{-2E(Y_j\widehat{f}_{n,M_i}(x_j)) + 2E(Y_{new,j})E\widehat{f}_{n,M_i}(x_j))\} \\
= \; & -\frac{2}{n}\sum_{j=1}^{n}COV(\widehat{f}_{n,M_i}(x_j), Y_j)
\end{aligned}
$$

Note that $\widehat{f}_{n,M_i}(x_j)$ is now an estimator and not the estimate. This means that we must consider the $Y's$ used in the estimator as stochastic variables. It is used that $E(Y_j) = E(Y_{new,j})$, $E(Y_j^2) = E(Y_{new,j}^2)$ and that $Y_{new,j}$ is independent of $\widehat{f}_{n,M_i}(x_j)$ since $Y_{new,j}$ is not used to estimate any of the parameters in $\widehat{f}_{n,M_i}(x_j)$.

It follows that an unbiased estimator $\widehat{E}_n[(Y_{new} - \hat{y}_{n,M_i,new})^2]$ of $E[(Y_{new} - \hat{y}_{n,M_i,new})^2]$ is given by

$$
\widehat{E}_n[(Y - \hat{y}_{n,M_i,new})^2] = \frac{1}{n}\sum_{j=1}^{n}(y_j - \widehat{f}_{n,M_i}(x_j))^2 + \frac{2}{n}\sum_{j=1}^{n}COV(\widehat{f}_{n,M_i}(x_j), Y_j) \tag{5.5}
$$

Equation (5.5) is an intuitive and informative result.[48] We see that the empirical plug-in estimate must be punished by $\frac{2}{n}\sum_{j=1}^{n}COV(\widehat{f}_{M_i}(x_j), Y_j)$ to be unbiased. The more influence each observation has on the associated prediction of a response with the same covaraite combination, the more the empirical plug-in estimate must be punished for sample bias. Although this does not clearly appear as a direct function of the dimension of the model, $|M_i|$, it is easy to argue that it is. Imagine, for instance, a very simple normal linear regression model with only the constant term. Since the MLE estimate of the constant in this situation only will be the average of the $y's$, each y will have small impact on the MLE. However, with no covariates the sum of the squared losses are likely to be larger. As we include covariates, each observation is likely to be more and more correlated the associated estimated value, increasing the need punish for sample bias.

Now, let $\hat{f}_{n,M_i} = (\hat{f}_{n,M_i}(x_1), ..., \hat{f}_{n,M_i}(x_n))^t$ be the vector of fitted values to the original data set. A particular insightful application of the result in equation (5.5) is when the vector $\hat{f}_{n,M_i}$is a linear combination of the Y vector where the $Y's$ are independently distributed with $VAR(Y) = \sigma^2$. More precisely, let

$$
\hat{f}_{n,M_i} = S_{M_i}Y
$$

[48] See also Hastie et al. (2009) p. 229 and Wasserman (2003) p. 219 for a discussion of this result.

where S_{M_i} is a $n \times n$ projection matrix. Let [i,i] indicate that we are in row and column i of a matrix, i.e. at the the diagonal. Then we have

$$\begin{aligned}
\sum_{j=1}^{n} COV(\widehat{f}_{n,M_i}(x_j), Y_j) &= \sum_{j=1}^{n} COV(\hat{f}_{n,M_i}, Y)[j,j] \\
&= \sum_{j=1}^{n} COV(S_{M_i} Y, Y)[j,j] \\
&= \sum_{j=1}^{n} S_{M_i} COV(Y,Y)[j,j] \\
&= \sum_{j=1}^{n} S_{M_i} \sigma^2 I_n[j,j] \\
&= \sigma^2 \sum_{j=1}^{n} S_{M_i}[j,j] \\
&= \sigma^2 Tr(S_{M_i})
\end{aligned}$$

Hence, we get that an unbiased estimator of $E[(Y_{new} - \hat{y}_{n,M_i,new})^2]$ in this particular situation is

$$\widehat{E}_n[(Y - \hat{y}_{n,M_i,new})^2] = \frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{f}_{n,M_i}(x_j))^2 + \frac{2}{n} \sigma^2 Tr(S_{M_i})$$

Consequently, the punishment for complexity is proportional to $Tr(S_{M_i})$. $Tr(S_{M_i})$ can be considered as the effective number of parameters[49] as will be motivated in more detail shortly. The relationship between $Tr(S_{M_i})$ and $|M_i|$ can be seen by assuming the normal linear regression model.

If we assume a normal linear regression model, the relationship between $\sum_{j=1}^{n} COV(\hat{f}_{n,M_i}(x_j), Y_j)$ and the dimension of the model $|M_i|$ becomes apparent. Assume that we fit the model M_i with p+1 covariates associated with parameters $\beta = (\beta_0, \beta_2, ..., \beta_p)^T$. Thus, we have that

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i = x_i^t \beta + \varepsilon_i$$

i=1,...n. Now we have that $\widehat{f}_{n,M_i}(x_j) = x_j^t \hat{\beta}$.

The system can also be written on matrix form

$$Y = X\beta + \varepsilon$$

[49] See Hastie et al. (2009) p. 232.

where X is the $n \times (p+1)$ covariate matrix, where the first column is consisting of 1's to take account for the constant term. The dimension of the model is $|M_i| = p+2$, taking into account that we have p+1 covariate parameters and the parameter σ^2.

From the standard regression literature[50], we know that the MLE parameters for β is

$$\hat{\beta}_n = (X^t X)^{-1} X^t y$$

Hence, we have

$$S_{M_i} = X\hat{\beta}_n = X(X^t X)^{-1} X^t$$

Using regular rules of trace operations we get that

$$\begin{aligned} Tr(S_{M_i}) &= Tr(X(X^t X)^{-1} X^t) \\ &= Tr(X^t X (X^t X)^{-1}) \\ &= Tr(I_{p+1}) \\ &= p+1 \end{aligned}$$

since X is a $n \times (p+1)$ matrix. Hence, we get that the unbiased estimate of the mean squared error is

$$\begin{aligned} \widehat{E}_n[(Y - \hat{y}_{n,M_i})^2] &= \frac{1}{n}\sum_{j=1}^{n}(y_j - x_j^t \hat{\beta}_{n,M_i})^2 + \frac{2}{n}\sigma^2(p+1) \\ &= \hat{\sigma}^2_{n,M_i} + \frac{2}{n}\sigma^2(p+1) \end{aligned}$$

This result is interesting for interpretation. We see that we punish model complexity increasingly in the number of covariates, p+1. The factor of the punishment is $\frac{2}{n}\sigma^2$. Not surprisingly the punishment is decreasing in n. Maybe less obvious is that the punishment is increasing in σ^2. A plausible explaination for this is that the more random variation, the more room for awarding spurious relationships, and the more the number of parameters should be punished.

Since we don't know σ^2, the result is not directly applicable for estimation. A possible alternative is to use an estimate of σ^2, but one should then preferably use one that is common for all models. An alternative is to use the $\hat{\sigma}^2_{M_w}$ where M_w is the widest of all possible models. In fact using $2\hat{\sigma}^2_{M_w}(p+1)$ as a general punishment term to the residual sum of squares of model M_i as a punishment for complexity is the basis for the Mallows C_p statistics used in model selection.[51]

If the real economic loss is a linear transformation of this squared error, we can simply use $a + b\widehat{E}_n[(Y - \hat{y}_{n,M_i})^2]$ to calculate the expected loss of a model M_i.

[50] See, for instance, Wasserman (2003) p. 217.

[51] See Wasserman (2003) p. 219.

We see that in the normal linear regression setting it is easy to straightforward calculate the bias correction term on a general level. For other models, the calculations soon become messy. Instead of calculating the correction term can be calculated by simulations or approximation techniques. Generally, it will be easier in these cases to use cross-validation as will be explained in more detail below.

5.3.3 Zero-one loss in two class prediction[52]

Assume now that we have a zero-one loss for the prediction error in a two class prediction.

$$L(Y_{new}, \hat{y}_{n,M_i,new}) = L_1 I(\hat{y}_{,n,M_i,new} = 0 \mid Y_{new} = 1)) + L_0 I(\hat{y}_{,n,M_i,new} = 1 \mid Y_{new} = 0)$$

Assume that we use a binomial GLM to predict class. Hence, a binomial regression model is used to estimate $\widehat{p}_{n,M_i}(x_j)$, the probability that Y_{new} belongs to class 1 given covariate combination x_j. If we knew the "true" probability $p(x_j)$, the expected loss (risk) would be minimized by a decision rule to choose class 1 if $p(x_j) > \frac{L_1}{L_1+L_2}$ and class 0 otherwise. The problem is that we don't know the true $p(x_j)$. Assume that we use the same principle, but instead rely on the estimate, and let $\hat{y}_{n,M_i,new} = 1$ if $\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2}$and zero else. Hence, for a given covariate combination x_j, we get

$$\begin{aligned} L(Y_{new}, \hat{y}_{n,M_i,new} \mid x_j) &= L_1 I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2} \mid Y_{new} = 1, x_j) \\ &\quad + L_0 I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2} \mid Y_{new} = 0, x_j) \end{aligned}$$

Since Y_{new} is independent of $\widehat{p}_{M_i}(x_j)$, which is estimated from observed data . This gives us

$$\begin{aligned} E[L(Y_{new}, \hat{y}_{n,M_i,new} \mid x_j)] &= L_1 E[I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2})] E I(Y_{new} = 1 \mid x_j)] \\ &\quad + L_0 E[I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2})] E I(Y_{new} = 0 \mid x_j)] \end{aligned}$$

We want an unbiased estimate of

$$E[L(Y_{new}, \hat{y}_{n,M_i,new})] = \frac{1}{n} \sum_{j=1}^{n} E[L(Y_{new}, \hat{y}_{n,M_i,new} \mid x_j)]$$

The plug-in empirical distribution estimate is

[52] This section is partially inspired by Friedman (1997).

$$E_n^*[L(Y_{new},\hat{y}_{n,M_i,new})] = \frac{1}{n}\sum_{j=1}^{n}\{L_1 I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2})I(Y_j=1) + L_0 I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2})I(Y_j=0)\}$$

To find the bias of the the plug-in empirical estimate, we use that

$$\begin{aligned}
&E[E_n^*[L(Y_{new},\hat{y}_{n,M_i,new})]] - E[L(Y_{new},\hat{y}_{n,M_i,new})] \\
&= E[\frac{1}{n}\sum_{j=1}^{n}\{L_1 I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2})I(Y_j=1) \\
&\quad + L_0 I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2})I(Y_j=0)\}] \\
&\quad - \frac{1}{n}\sum_{j=1}^{n} E[L(Y_{new},\hat{y}_{n,M_i} \mid x_j)] \\
&= E[\frac{1}{n}\sum_{j=1}^{n}\{L_1 I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2})I(Y_j=1) \\
&\quad + L_0 I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2})I(Y_j=0)\}] \\
&\quad - \frac{1}{n}\sum_{j=1}^{n}\{L_1 E[I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2})]EI(Y_{new}=1 \mid x_j)] \\
&\quad + L_0 E[I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2})]EI(Y_{new}=0 \mid x_j)]\} \\
&= \frac{1}{n}\sum_{j=1}^{n}\{L_1 COV(I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2}), I(Y_j=1)) \\
&\quad + L_0 COV(I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2}), I(Y_j=0))\} \qquad (5.6)
\end{aligned}$$

Hence, the last term ove equation (5.6) is the over-optimism of the empirical estimate to be used as a bias correction term. This is an interesting result. Since $I(Y_j=1) = 1 - I(Y_j=0)$, we can rewrite the bias correction expression as

$$-\frac{1}{n}\sum_{j=1}^{n}\{L_1 COV(I(\widehat{p}_{n,M_i}(x_j) \leq \frac{L_1}{L_1+L_2}), I(Y_j=0)) + L_0 COV(I(\widehat{p}_{n,M_i}(x_j) > \frac{L_1}{L_1+L_2}), I(Y_j=1))\} \quad (5.7)$$

As with the mean square error, the plug-in empirical estimate is biased upwards and must be corrected. The correction is higher the higher the influence each observation has on the outcome of the fitted classification as we explained for the mean squared error loss above.

In the special case of $L_1 = L_2 = 1$, the equation (5.7) reduces to

$$-\frac{1}{n}\sum_{j=1}^{n}\{COV(I(\widehat{p}_{n,M_i}(x_j) \leq 0.5), I(Y_j = 0) + COV(I((\widehat{p}_{n,M_i}(x_j) > 0.5), I(Y_j = 1))\}$$

$$= -\frac{2}{n}\sum_{j=1}^{n} COV(I(\widehat{p}_{n,M_i}(x_j) \leq 0.5), I(Y_j = 0))$$

Hence in this case, we have that an unbiased estimator for $E[L(Y_{new}, \hat{y}_{n,M_i,new})]$ is

$$\begin{aligned}\widehat{E}[L(Y_{new}, \hat{Y}_{new,M_i})] &= \frac{1}{n}\sum_{j=1}^{n}\{I(\widehat{p}_{n,M_i}(x_j) \leq 0.5)I(Y_j = 1) + I(\widehat{p}_{n,M_i}(x_j) > 0.5)I(Y_j = 0)\} \\ &\quad + \frac{2}{n}\sum_{j=1}^{n} COV(I(\widehat{p}_{n,M_i}(x_j) \leq 0.5), I(Y_j = 0))\end{aligned}$$

By comparing to equation (5.5) above, we see now clearly that the bias correction principle corresponds to the squared error loss case. As for the squared error loss, the direct calculation of

$$\sum_{j=1}^{n} COV(I(\widehat{p}_{n,M_i}(x_j) \leq 0.5), I(Y_j = 0))$$

becomes messy. Cross-validation, which we will return to below, is a good alternative.

5.3.4 AIC in an another loss estimation perspective

We can also fit the AIC into the $L(\mu, \hat{\mu}_{n,M_i})$ framework. This might make the AIC more intuitive and less "mystical" by not associating it with K-L distance. Assume that the utility of using a model is the expected log likelihood of the new response we want to predict, Y_{new}. The higher the expected log likelihood of the prediction the higher the utility. Transforming this into a loss framework would be to let the loss be the minus expected log likelihood, which we want as small as possible. In other words, we have

$$L(\mu, \hat{\mu}_{n,M_i}) = -\,\mathrm{E}(\log f(Y_{new}; \hat{\theta}_{n,M_i}))$$

The expected loss is

$$E[L(\mu, \hat{\mu}_{n,M_i})] = E[E(\text{ -log } f(Y_{new}; \hat{\theta}_{n,M_i}))]$$

If we go back to section 3.3.1 we see that this is the same that we want to estimate in the derivation of AIC. The outer expectation is with respect to the MLE which is a random variable.

Since we want to minimize over the equal weight "average" of all covariates, we have

$$E[\text{ -log } f(Y_{new};\hat{\theta}_{n,M_i})] = \frac{1}{n}\sum_{j=1}^{n} E[\text{ -log } f(Y_{new,j} \mid x_j;\hat{\theta}_{M_i})$$

By using the result stated in section 3.1.1, we find that an approximately unbiased estimate of $E[\text{ -log } f(Y_{new};\hat{\theta}_{n,M_i})]$ is

$$\begin{aligned}\widehat{E}_n[\text{ -log } f(Y_{new};\hat{\theta}_{n,M_i})] &= \frac{1}{n}\sum_{j=1}^{n} \text{-log } f(Y_j \mid x_j;\hat{\theta}_{M_i}) + \frac{|M_i|}{n} \\ &= -\frac{AIC(M_i)}{2n}\end{aligned}$$

We have now seen that AIC (divided by 2n) can be considered as an estimate of the expected loss of prediction when using a model where the loss is the expected negative log likelihood of using a model. On the contrary to squared error loss with normal linear regression described above, the simple bias correction term is independent of the probability density used. The same parameter complexity punishment is used independent on the actual distribution function used for modeling. This makes AIC an easy accessible method in the direct estimation of expected loss.

Using expected minus log likelihood as loss function can appear rather odd as a loss function when the purpose of the loss function is to reflect some real economic value. However, it can be argued that at least a linear transformation of the minus log likelihood of can be used to approximate the real economic loss of using a model.

5.3.5 The usefulness of direct expected loss estimation

As we have seen above the direct unbiased estimation of expected loss of using a model by the plug-in empirical distribution estimate and an appropriate bias correction is possible for some loss functions. Hence, it is at least in theory an applicable method to calculate the information value of covariates in real economic contexts.

However, it seems that even for simple loss functions the calculation of a proper bias correction term when using the plug-in empirical distribution estimate becomes messy. We cannot ignore such term either, because a more complex model will always reduce the estimated loss. Based on the analysis above it seems like direct estimation is feasible and can be recommended in at least two cases. If the expected loss of using a model is the mean squared error and that we choose covariates in a normal regression setting, the direct estimation of an unbiased estimate of the expected loss associated with a model is possible. The same method will also work when the expected loss is a linear transformation of the mean squared error.

The other case is if the loss is the minus log likelihood associated with a model, or a linear transformation of the linear log likelihood. In that case the penalty term for adding more parameter (and hence) covariate is independent of the probability distribution used for estimation.

5.4 A cross-validation approach to expected loss estimation in prediction settings

5.4.1 Estimating expected loss by cross-validation

Cross-validation (CV) is a technique for estimation primary developed to validate models, but can also be used as a method of estimation as such, in particular as a method for expected loss estimation for model selection.[53]

Assume that a sample of n is randomly divided into K roughly equal sized partitions, n_k, k=1,...,K . Let $k(j)$ be the indexing function that indicates the partition observation j is allocated to. Furthermore, let $\hat{\mu}^{-k}_{n-n_k,M_i}$ be the MLE based predictor of μ under model M_i computed without the k-partition of the data. The k-fold CV estimated expected loss under model M_i can then be written as

$$\widehat{E}_n^{CV(k)}[L(\mu,\hat{\mu}_{n,M_i})] = \frac{1}{n}\sum_{j=1}^{n} L(\mu_j,\hat{\mu}^{-k(j)}_{n-n_k,M_i} \mid x_j)$$

where x_j is the covariate vector associated with observation j.

We see that if we simply are going to predict the response variable Y_{new} in a regression, the equation reduces to

$$\widehat{E}_n^{CV(k)}[L(Y_{new},\hat{y}_{n,M_i,new})] = \frac{1}{n}\sum_{j=1}^{n} L(y_j,\hat{y}^{-k(j)}_{n-n_k,M_i,j} \mid x_j)$$

A special case is the case where k=n. In this case each partiation consists the full data set minus one observation. This is known as n-fold cross-validation and leave-one-out cross-validation. In this case the estimation is based on all observations other than the one to be predicted. The leave one out cross-validation estimation of the expected loss can simply be written as

$$\widehat{E}_n^{CV(n)}[L(\mu,\hat{\mu}_{M_i})] = \frac{1}{n}\sum_{j=1}^{n} L(\mu_j,\hat{\mu}^{-j}_{n-1,M_i} \mid x_j)$$

In the setting where our focus is to predict a new Y_{new} this becomes

$$\widehat{E}_n^{CV(n)}[L(Y_{new},\hat{y}_{n,M_i,new})] = \frac{1}{n}\sum_{j=1}^{n} L(y_j,\hat{y}^{-j}_{n-1,M_i,j} \mid x_j)$$

[53] A survey of the use of cross-validation in model selection is provided by Arlotte and Celisse (2010).

In the next subsections we will first say more about the properties of the CV estimate of the expected loss. After that we will address some problems associated with CV expected loss estimation, and briefly discuss jackknife and bootstrapping, which are related methods to CV.

In the special case of a linear model where we have $\hat{f}_{M_i} = S_{M_i}Y$ and $E[L(\mu, \hat{\mu}_{M_i})] = E[(Y_{new} - \hat{y}_{n,M_i,new})^2]$, we have that[54]

$$\begin{aligned}\widehat{E}_n^{CV(n)}[L(Y_{new}, \hat{y}_{n,M_i,new})] &= \frac{1}{n}\sum_{j=1}^{n}(Y_j - \hat{f}_{n,M_i}^{-j}(x_j))^2 \\ &= \frac{1}{n}\sum_{j=1}^{n}\left[\frac{Y_j - \hat{f}_{n,M_i}(x_j)}{1 - S_{M_i}[j,j]}\right]^2\end{aligned}$$

where $\hat{f}_{n,M_i} = (\hat{f}_{n,M_i}(x_1), ..., \hat{f}_{n,M_i}(x_n))^t$ and $S_{M_i}[j,j]$ is the j'th diagonal element in S_{M_i}. In this particular case we don't need to find the MLE for all the n partitions of the data set.

5.4.2 What do we estimate when using cross validation?

A first natural question is what the CV estimate of the expected loss really is, both as such and asymptotically. We would like to note that statistical properties of the CV-estimate is a major topic subject to extensive contemporary research.[55] We have no ambition to give a comprehensive discussion here, nor do we have any ambition to touch upon the research frontiers of this topic. We will simply discuss some of the topics of most relevance for this study.

We recall from the previous chapters that we simply postulated that we were interested in the estimation of expected loss averaged over the covariates in the sample. We gave each observation in the sample equal weight as follows:

$$E[L(\mu, \hat{\mu}_{n,M_i})] = \frac{1}{n}\sum_{j=1}^{n}E[L(\mu, \hat{\mu}_{n,M_i} \mid x_j)]$$

Hence, we consider the training covariates as non-random and having equal weight. We find the expected loss with respect to variations in responses for these non-random covariates. By this we estimate the in-sample expected loss. When we searched for the bias using the plug-in empirical distribution estimate above, we corrected for this in-sample bias. Hence, the question is if we estimate the same in-sample expected loss when using leave-one-out cross-validation. We will shed some light on this question.

For a given covariate combination, we have that

$$E[L(\mu_{new}, \hat{\mu}_{n-1,M_i}^{-j} \mid x_j)] = E[L(\mu_j, \hat{\mu}_{n-1,M_i}^{-j} \mid x_j)$$

[54] See for instance Hastie et al. (2009) p. 244 or Claeskens and Hjort (2008) p. 55.

[55] See Arlotte and Celisse (2010) for a survey article.

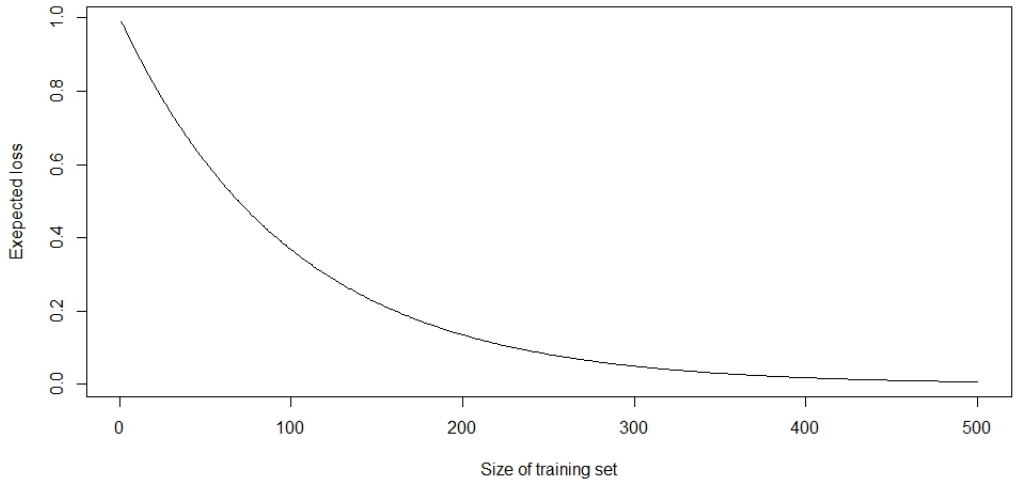

Figure 5.2: The learning curve of data

where μ_j is the μ associated with observation j. The reason is that data used in the estimation $\hat{\mu}^{-j}_{j,M_i}$ is not used compute μ_j. More precisely, the observations $(y_1,...y_{j-1},y_{j+1},...,y_n)$ do not contain the y_j used to compute μ_j, and the observations are independent by assumption. For instance, if our focus is to predict Y_{new} for covariate combination x_j, we have that

$$E[L(Y_{new},\hat{y}^{-j}_{n-1,M_i,new} \mid x_j)] = E[L(Y_j,\hat{y}^{-j}_{n-1,M_i,new} \mid x_j)]$$

There are however several reasons why the overall estimate $E[L(\mu,\hat{\mu}_{n,M_i})]$ by cross-validation is biased or improper as an estimate for $\frac{1}{n}\sum_{j=1}^{n} E[L(\mu,\hat{\mu}_{n,M_i} \mid x_j)]$.

One source of bias is the learning curve bias.[56] Estimating parameters on the basis of $n-n_k$ gives less precise estimates than estimating on the basis of the full n, with the result that the estimated loss is exaggerated. The result is that the expected loss is over-estimated.[57] The impact of over-estimation is depending on where we are on the learning curve of the data. The idea is that the marginal reduction is expected loss is decreasing as the training set increases. This is illustrated by a hypothetical learning curve in Figure 5.2. As we see more data reduces our expected loss but, more data have more impact on the expected loss when data is already scarce. The learning curve effect has naturally most impact i k-fold cross-validation when K<<n. In the leave-one-out setting it is reasonable to assume that the learning

[56] See for instance Hastie et al. (2009) p. 243 and Arlotte and Celisse (2010).

[57] see Hastie et al. (2009) p. 243.

curve effect is not a substantial source of bias, as long as n is not very low.

One might ask what the point is to use K<<n if this increases the bias. What is the benefit? The optimal size of K is subject to on-going research and a discussed topic.[58] One benefit of having K<<n is that the variance of the estimate might be less.[59] We will not go into the variance issue in this study, but comment on variance issues in the concluding remarks in Chapter 7.

Furthermore, in general, we cannot now simply postulate that the covariates are non-random and give them equal weight and expect that CV will provide us with this estimate. We will get an estimate for the out-of-sample expected loss.[60] We see that if the covariate combinations really are independent and are equally probable, we get

$$E_X E[L(\mu, \hat{\mu}_{n,M_i})] = \frac{1}{n}\sum_{j=1}^{n} E[L(\mu, \hat{\mu}_{n,M_i} \mid x_j)]$$

and what we get is an unbiased estimate of the in-sample expected loss. However, if there is an underlying distribution associated with the covariates, the cross-validation will capture it. The practical consequence of this is that in addition to the in-sample bias of the training data, one get an out-of-sample bias due to the fact that that we do the estimation of one training set among many possible training sets, including variations in the covariates. The cross-validation estimate gives us an estimate correcting for the out-of-sample bias, whether we want it or not. A consquence of this is that the estimated expected loss is likely to be larger than when out-of-sample bias is assumed away.

If we go back to Chapter 3, our choice of the in-sample expected loss as an estimation objective was a practical one. We did not want add the complication of allowing for covariate distribution. However, what we really want is the estimate of the expected loss of using a model. So in some sense it could be argued that out-of-sample expected loss is a better one. For large data the two measurements will converge as described in section 3.3.3.

5.4.3 Computational issues when using cross validation in estimation

An obvious problem with cross-validation is the computation intensity required to obtain the estimates. In leave-one-out cross-validation, we will need to make n estimations of the maximum likelihood parameters. For a large amount of data and when complicated numerical methods must be used in estimation, the estimation can be time consuming. When this in addition must be done for various candidate models, leave-one-out cross-validations methods might not seem practical.

Although, computational issues may not seem to impose constraints in many academic settings, computing constraints is obvious in many practical settings. In financial settings, where decisions must be taken within seconds, computational issues will soon impose real constraints. Another example is

[58] See Hastie et al. (2009) p. 243 and Arlotte and Celisse (2010).

[59] Hastie et al. (2009) p. 243.

[60] See Hastie et al. (2009) p. 254.

weather forecasting. It does not help much to find a good cost-efficient model to predict the weather condition in one hour, if it takes one hour to find this model.

We don't believe however, that computational constraints in most cases is a big issue in our context. Recall, that for many purposes, even the financial and weather settings described above, we might have have good time to find the cost-efficient model before we are under crucial time-constraints. For most cases our cross-validation estimation can run overnight without problems. The real time constraint will then be more related to the time it takes to apply the chosen model.

Furthermore, we see that computational constraints might be an issue when there are hundreds of measurement to choose between to choose among to use as covarites, big data sets, and we are employing methods such as Lasso regression. However, for this study, we have assumed that we operate in a context with not to many alternative models, and we assume that standard GLM models are used. For such models fast running MLE estimation algorithms have been developed. Hence, in our context we don't consider computational resources as a main constraint in using cross-validation to estimate the expected loss associated with models.

5.4.4 The usefulness of cross-validation in expected loss estimation

In our setting cross-validation seems like a general and applicable method to estimate the expected loss associated with a model used to predict a certain focus. If we are outside the domain of some standard loss functions in some specific regression settings, cross-validation seems preferable to direct estimation.

One problem of cross-validation is that it prevents the full use of all available data in the estimation. Depending on where we are on the learning curve of data, this might be a problem. However, this is not a major issue i n-fold cross-validation as we use nearly all data. Furthermore, the use of the whole data-set and cross-validation can be combined by the jackknife method which will be briefly discussed below.

Cross-validation requires computational power and skills. However, in the simple setting we use here, narrowing our analysis to GLM models, this doesn't seem to constitute a major obstacle to the use of cross-validation in the estimation of the expected loss associated with various models.

5.4.5 Jackknife estimation

A problem of cross-validation estimation of expected loss is that we don't utilize the full data which may result in an overestimation bias as described above. Although the impact of this overestimation is usually small when we use n-fold cross-validation (leave-one-out), it is at least a problem in principle for cross-validation. To utilize the full data set and still use the bias correcting benefit of leave one-out-cross validation, we could use the cross-validation for exact that purpose: to correct for for bias. This is done in Jackknife estimation.[61] We will just briefly discuss the idea here to illustrate how the Jackknife can be used to estimate expected loss functions.

[61] See for instance Knight (2000) p. 226.

Assume that we want to estimate α. Le $\hat{\alpha}_n$ be the estimate of α using the full data set of n observations. Furthermore let $\hat{\alpha}_{n-1}^{-j}$, be the estimate of α, with observation j, left out. From the previous sections we know that

$$\hat{\alpha}_n^{CV(n)} = \frac{1}{n}\sum_{j=1}^{n} \hat{\alpha}_{n-1}^{-j}$$

$\hat{\alpha}_n^{CV(n)}$ is the leave-one out cross validation estimator of α. Now, let the Jackknife estimator be given by

$$\hat{\alpha}_n^{JK} = n\hat{\alpha} - (n-1)\hat{\alpha}_n^{CV(n)}$$

The Jackknife estimator is based on an assumption that we can write

$$E(\hat{\alpha}_n) = \alpha + \sum_{j=1}^{z} \frac{a_j(\alpha)}{n}$$

In general, if z=1, the Jackknife estimator is unbiased. If z>1 then the jackknife estimator will still be biased, but less biased than $\hat{\alpha}_n$ for n sufficiently large.

Hence, if we let $\alpha = E[L(\mu, \hat{\mu}_{n,M_i})]$, we can use the jackknife to estimate the estimated loss as an alternative to the cross-validation estimator. We will not pursue this issue here, but note that the Jackknife can improve the cross-validation estimator.

5.4.6 Bootstrapping

If you try to lift yourself by pulling your own boots you will not get high. Luckily, bootstrapping in statistics gets you much further than that. Bootstrapping is a technique involving resampling of data to find properties of estimates. Simplified, we draw different samples by sampling a random selection of data from a data-set and use this data set for estimation. By this methods you can get many estimates based on the same data.

Bootstrapping can complement cross-validation loss estimation and other methods of loss estimation by allowing us to get a more or less good picture of the variance and other statistical properties of the estimator. We will not pursue bootstrapping further in this study, but note that it is a useful method to complement the methods used in this study. We will, however, briefly discuss variance estimation in the concluding remarks in Chapter 7.

5.5 A FIC-inspired approach to expected loss estimation

Above, we assumed that our focus was to predict a new response for which we have observations (or a function of such response only involving observable variables). Often, we want a to use the model not to predict a new response, but to estimate a parameter, for instance the 95% percentile of the response variable. The focused information criterion (FIC) is developed as a powerful tool to choose among

models that aims at minimzing the mean squared error of a focus estimate. FIC was described above in Section 3.3.2.

As a starting point the use of FIC might seem unsuitable in a regression setting when choosing model for an unknown combination of covariates, since FIC assumes that we want to choose model for a specific combination of covariates as a focus. However, averaging over FIC values or using the "averaged FIC", AFIC, is suitable for this situation, because it allows us to choose a model based on an average of focuses in a regression setting. Below we will continue the description of FIC started in section 3.3.2, and explain in more detail how averaging FIC values or using AFIC can be suitable for our situation where we want to calculate the information value of gathering covariates not yet known. After this we will explore how FIC can be of use to calculate the information value of gathering covariates, even if our loss function is not the squared loss.

5.5.1 From FIC to AFIC

As described in section 3.3.2, the FIC is based on an assumption that the $Y's$ are distributed according to the following distribution

$$f_n(y) = f(y \mid \theta_0, \gamma_0 + \delta/\sqrt{n})$$

This distribution applies well to generalized regression settings where a typical model section issue is covariate selection. Then we will typically have

$$\gamma_0 + \delta/\sqrt{n} = (\beta_0, ..., \beta_p)^t = \beta$$

A typical focus μ in a regression setting could be the expected Y for a particular covariate combination x_0. In the normal linear regression context, this would give us

$$\mu_{true} = \beta_0 + \beta_1 x_{01} + ... + \beta_p x_{0p}$$

If we assume that $M_{\{0klm\}}$ is a sub-model consisting of the constant term and the covariates k,l,m we can write

$$\hat{\mu}_{n,M_{\{0klm\}}} = \hat{\beta}_{0,n,M_{\{0klm\}}} + \hat{\beta}_{k,n,M_{\{0klm\}},k} x_{0k} + \hat{\beta}_{l,n,M_{\{0klm\}}} x_{0l} + \hat{\beta}_{m,n,M_{\{0klm\}}} x_{0m}$$

In the regression setting we have in the limit based on the limit distribution, using the FIC framework described in section 3.3.2:

$$VAR(\hat{\mu}_{n,M_i}(x_0)) = \frac{1}{n}(\tau_0^2 + \omega(x_0)^t Q^0_{M_i} \omega(x_0))$$

and

$$\begin{aligned} BIAS(\hat{\mu}_{n,M_i}(x_0)) &= E(\hat{\mu}_{M_i}(x_0)) - \mu_{true}(x_0) \\ &= \frac{1}{\sqrt{n}}(\omega(x_0)^t(\delta - G_{M_i}\delta)) \end{aligned}$$

This gives us the following limit

$$\begin{aligned} MSE(\hat{\mu}_{n,M_i}(x_0)) &= E[(\hat{\mu}_{n,M_i}(x_0) - \mu_{true}(x_0))^2] \\ &= \frac{1}{n}(\tau_0^2 + \omega(x_0)^t Q^0_{M_i} \omega(x_0) + \omega(x_0)^t (I_q - G_{M_i}) \delta \delta^t (I_q - G_{M_i})^t \omega(x_0)) \end{aligned}$$

with notation as described in section 3.3.2. Based on this, the estimated MSE for $\hat{\mu}_{n,M_i}(x_0)$ becomes

$$\begin{aligned} \widehat{MSE}_n(\hat{\mu}_{n,M_i}(x_0)) &= \frac{1}{n}(\hat{\tau}^2_{0,n} + \hat{\omega}_n(x_0)^t \hat{Q}^0_{n,M_i} \hat{\omega}_n(x_0) \\ &\quad + \hat{\omega}_n(x_0)^t (I_q - \hat{G}_{n,M_i})(\hat{\delta}_{n,wide} \hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \hat{G}_{n,M_i})^t \hat{\omega}_n(x_0) \end{aligned}$$

FIC$(\hat{\mu}_{n,M_i}(x_0))$ is obtained by multiplying $\widehat{MSE}_n(\hat{\mu}_{n,M_i}(x_0))$ by n. As described in section 3.2.2, a negative estimated bias squared term can be avoided by truncation. This is obtained by replacing $\hat{\omega}_n(x_0)^t (I_q - \hat{G}_{n,M_i})(\hat{\delta}_{n,wide} \hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \hat{G}_{n,M_i})^t \hat{\omega}_n(x_0)$ by

$$max\{0, \hat{\omega}_n(x_0)^t (I_q - \hat{G}_{n,M_i})(\hat{\delta}_{n,wide} \hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \hat{G}_{n,M_i})^t \hat{\omega}_n(x_0)\}$$

For our purpose, which is to calculate the information value of gathering covariates, we don't know x_0. The question is which of the components of x_0 to gather. The approach to this issue in this study has been to estimate the expected loss of a model based on an average of the covariates in the sample. The direct approach to finding this average would be to just take the average $\widehat{MSE}(\hat{\mu}_{n,M_i}(x))$ over all the $x's$ in the sample. This is to let

$$\widehat{MSE}_n(\hat{\mu}_{n,M_i}) = \frac{1}{n} \sum_{j=1}^{n} \widehat{MSE}_n(\hat{\mu}_{n,M_i}(x_j))$$

or correspondingly, by using FIC

$$FIC(\hat{\mu}_{n,M_i})) = \frac{1}{n} \sum_{j=1}^{n} FIC(\hat{\mu}_{n,M_i}(x_j))$$

As we see, under this approach we have to explicitly find the estimated MSE or the corresponding FIC for each covariate combination.

An alternative is to approximate the distribution for the averaged focus, which allows us to estimate the average MSE/FIC directly. This is the idea behind AFIC (Averaged Focused Information Criterion) as described in more detail in Claeskens and Hjort (2008) Chapter 6.9. AFIC is a general measurement to find the FIC value for a weighted average of focuses. A special case of this is to approximate the FIC

averaged over the covariates in the sample in a regression setting. Recalling equation (3.14), we have for each individual covariate combination x:

$$\sqrt{n}(\hat{\mu}_{n,M_i}(x) - \mu_{true}(x)) \xrightarrow{d} \Lambda_{M_i}(x) = \Lambda_0(x) + \omega(x)^t(\delta - G_{M_i}D)$$

We can now consider the loss function

$$\begin{aligned} L_n(\hat{\mu}_{M_i}) &= n\int (\hat{\mu}_{n,M_i}(x) - \mu_{true}(x))^2 dW(x) \\ &= \int nMSE(\hat{\mu}_{M_i}(x))dW(x) \end{aligned} \tag{5.8}$$

Where $W(x)$ is the distribution of weights for the x's. If we assume that W converges or is simply fixed as in our case (to 1/n), we have under mild conditions[62]

$$L_n(\hat{\mu}_{n,M_i}) \xrightarrow{d} \int \Lambda_{M_i}(x)^2 dW(x)$$

The average expected loss (i.e risk) over the weight function W, $E[L_n(M_i)]$ will then converge to

$$E[\int \Lambda_{M_i}(x)^2 dW(x)] = \int E[\Lambda_{M_i}(x)^2]dW(x)$$

Using the results discussed in section 3.3.2, we find that

$$\begin{aligned} E[\Lambda_{M_i}(x)^2] &= \tau_0(x)^2 + \omega(x)^t Q^0_{M_i}\omega(x) + \omega(x)^t[(I_q - G_{M_i})\delta\delta^t(I_q - G_{M_i})^t]\omega(x) \\ &= \tau_0(x)^2 + Tr(Q^0_{M_i}\omega(x)\omega(x)^t) \\ &\quad + Tr((I_q - G_{M_i})\delta\delta^t(I_q - G_{M_i})^t\omega(x)\omega(x)^t) \end{aligned}$$

Now, let

$$A_0 = \int \tau_0(x)^2 dW(x)$$

$$A = \int \omega(x)\omega(x)^t dW(x)$$

Then we can write

$$\begin{aligned} \int E[\Lambda_{M_i}(x)^2]dW(x) &= A_0 + Tr(Q^0_{M_i}A) \\ &\quad + Tr((I_q - G_{M_i})\delta\delta^t(I_q - G_{M_i})^t A) \end{aligned}$$

[62] See Claeskens and Hjort (2008) p. 180.

Hence, in the limit we will have that

$$E[L_n(M_i)] = A_0 + Tr((I_q - G_{M_i})\delta\delta^t(I_q - G_{M_i})^t A) + Tr(Q^0_{M_i}A)$$

In Claeskens and Hjort (2008)[63], A_0 is ignored as it is a common element in all models. For model selection this work fine, as ignoring A_0 does not affect model choice. A_0 can also be ignored if we are interested in the information value of a parameter since it is then the difference in MSE between models we are interested in, and common terms will be canceled out. This cannot be done, however, if we are interested in the nominal value of the mean squared error. Since we, in this study, are primarily interested in calculating the nominal value of the MSE and other loss functions, we cannot ignore A_0. Using the proper estimates provides us with the following version of AFIC

$$AFICM(\hat{\mu}_{n,M_i}) = \widehat{A}_{0,n} + max\{0, \mathrm{Tr}((I_q - \widehat{G}_{n,M_i})(\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \widehat{G}_{n,M_i})^t\widehat{A}_n)\} + Tr(\widehat{Q}^0_{n,M_i}\widehat{A}_n)$$

For comparison, the AFIC in Claeskens and Hjort (2008) is

$$AFIC(\hat{\mu}_{M_i}) = max\{0, Tr((I_q - \widehat{G}_{n,M_i})(\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \widehat{G}_{n,M_i})^t\widehat{A}_n)\} + Tr(\widehat{Q}^0_{n,M_i}\widehat{A}_n)$$

As Claeskens and Hjort (2008)[64], we use a truncated version of the bias squared preventing a negative bias squared.

We are still not finnished, however. We are interested in an estimate for the average MSE. By inspecting equation (5.8), we see that this can be obtain by dividing AFICM by n. Hence,

$$\begin{aligned}\widehat{MSE}(\hat{\mu}_{n,M_i}) &= \frac{AFICM(\hat{\mu}_{n,M_i})}{n} \\ &= \frac{1}{n}(\widehat{A}_{0,n} + max\{0, Tr((I_q - \widehat{G}_{n,M_i})(\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}_n)(I_q - \widehat{G}_{n,M_i})^t\widehat{A})\} + Tr(\widehat{Q}^0_{n,M_i}\widehat{A}_n))\end{aligned}$$

5.5.2 Applying FIC to more general loss functions

Now assume that we have a general loss-function of the type

$$L(\mu_{true}, \hat{\mu}_{n,M_i}) = L(\hat{\mu}_{n,M_i} - \mu_{true}) \tag{5.9}$$

where we assume that $L(0) = 0$ and that L is smooth, positive and increasing in the error, i.e $L'(u) > 0$ for $u > 0$ and $L'(u) < 0$ for $u < 0$, which necessarily must mean that $L'(0) = 0$.

A second order Taylor development of $L(u)$ around u_0 gives us

[63] Claeskens and Hjort (2008) p. 180.
[64] Claeskens and Hjort (2008) p. 181.

$$L(u) = L(u_0) + L'(u_0)(u-u_0) + \frac{1}{2}L''(u_0)(u-u_0)^2 + \frac{1}{6}L'''(u^*)(u-u_0)^3$$

For some u^* between u and u_0.

For u close to u_0, we have

$$L(u) \approx L(u_0) + L'(u_0)(u-u_0) + \frac{1}{2}L''(u_0)(u-u_0)^2$$

Taking expectation on both sides gives

$$E[L(u)] \approx L(u_0) + L'(u_0)E(u-u_0) + \frac{1}{2}L''(u_0)E[(u-u_0)^2]$$

Now, let $u = \hat{\mu}_{M_i} - \mu_{true}$

$$E[L(\hat{\mu}_{n,M_i} - \mu_{true})] \approx L(u_0) + L'(u_0)E(\hat{\mu}_{n,M_i} - \mu_{true} - u_0) + \frac{1}{2}L''(u_0)E[(\hat{\mu}_{n,M_i} - \mu_{true} - u_0)^2]$$

First, let $u_0 = 0$. This gives us

$$\begin{aligned} E[L(\hat{\mu}_{n,M_i} - \mu_{true})] &\approx L(0) + L'(0)E(\hat{\mu}_{n,M_i} - \mu_{true}) + \frac{1}{2}L''(0)E[(\hat{\mu}_{n,M_i} - \mu_{true})^2] \\ &= \frac{1}{2}L''(0)E[(\hat{\mu}_{n,M_i} - \mu_{true})^2] \\ &= \frac{1}{2}L''(0)MSE(\hat{\mu}_{n,M_i}) \end{aligned}$$

An estimate for $E[L(\hat{\mu}_{M_i} - \mu_{true})]$ is then

$$\widehat{E}_n[L(\hat{\mu}_{n,M_i} - \mu_{true})] = \frac{1}{2}L''(0)\widehat{MSE}(\hat{\mu}_{n,M_i})$$

Using the results from FIC analysis to estimate the expected loss gives us

$$\widehat{E}_n[L(\hat{\mu}_{n,M_i} - \mu_{true})] = \frac{1}{2}L''(0)\frac{FIC(\hat{\mu}_{n,M_i})}{n}$$

We see that FIC can be used as an approximation to estimate the expected loss associated with a general loss function of the type given in (5.9). We see, however, that the approximation above is not likely to be very good as long as $\hat{\mu}_{M_i} - \mu_{true}$ is not close to zero. Furthermore we will not be able to capture the asymmetry in loss functions, for instance if we use a LINEX loss function.

We can, however, improve the approximation by a clever selection of u_0. If we had such u_0 we would get

$$\begin{aligned}
E[L(\hat{\mu}_{n,M_i} - \mu_{true})] &\approx L(u_0) + L'(u_0)E(\hat{\mu}_{n,M_i} - \mu_{true} - u_0) + \frac{1}{2}L''(u_0)E[(\hat{\mu}_{n,M_i} - \mu_{true} - u_0)^2] \\
&= L(u_0) + L'(u_0)(E(\hat{\mu}_{n,M_i} - \mu_{true}) - u_0) - L''(u_0)u_0E(\hat{\mu}_{n,M_i} - \mu_{true}) \\
&\quad + \frac{1}{2}L''(u_0)E[(\hat{\mu}_{n,M_i} - \mu_{true})^2] + \frac{1}{2}u_0^2L''(u_0) \\
&= L(x_0) - u_0L'(u_0) + (L'(u_0) - u_0L''(u_0))E(\hat{\mu}_{n,M_i} - \mu_{true}) \\
&\quad + \frac{1}{2}L''(u_0)E[(\hat{\mu}_{n,M_i} - \mu_{true})^2] + \frac{1}{2}u_0^2L''(u_0) \\
&= L(u_0) - u_0L'(u_0) + \frac{1}{2}u_0^2L''(u_0) + (L'(u_0) - u_0L''(u_0))BIAS(\hat{\mu}_{n,M_i}) \\
&\quad + \frac{1}{2}L''(u_0)MSE(\hat{\mu}_{n,M_i}) \qquad (5.10)
\end{aligned}$$

This gives us the following estimate for $E[L(\hat{\mu}_{n,M_i} - \mu_{true})]$ where we can use the results from the FIC analysis as estimates for bias and variance:

$$\begin{aligned}
\widehat{E}_n[L(\hat{\mu}_{n,M_i} - \mu_{true})] &= L(u_0) - u_0L'(u_0) + u_0^2L''(u_0) + (L'(u_0) - u_0L''(u_0))\widehat{BIAS}_n(\hat{\mu}_{n,M_i}) \\
&\quad + \frac{1}{2}L''(u_0)\widehat{MSE}_n(\hat{\mu}_{n,M_i}) \\
&= L(u_0) - u_0L'(u_0) + \frac{1}{2}u_0^2L''(u_0) + (L'(u_0) - u_0L''(u_0))\widehat{BIAS}_n(\hat{\mu}_{n,M_i}) \\
&\quad + \frac{1}{2}L''(u_0)\frac{FIC(\hat{\mu}_{n,M_i})}{n}
\end{aligned}$$

As we saw above, the FIC framework provides us with both an estimator of BIAS and MSE. The challenge is to find an u_0 close to $\hat{\mu}_{n,M_i} - \mu_{true}$. Since, $\hat{\mu}_{n,wide}$ is unbiased, a good candidate is likely to be

$$u_0 = \hat{\mu}_{n,M_i} - \hat{\mu}_{n,wide}$$

However, when doing this some cautionary notes are due. First observe that u_0, a priori, is stochastic. Hence, we cannot in principle take it out of the expectation as we do in equation (5.10). Still, replacing some measurement with its estimate is quite usual in statistics so we don't feel too uncomfortable about it. For instance, MLE are used as plug-ins when applying the delta-theorem for some unknown parameters, and it is done in the application of FIC. As $\hat{\mu}_{n,wide}$ is unbiased for μ_{true} the problem of doing so boils down to the variance of $\hat{\mu}_{n,M_i} - \hat{\mu}_{n,wide}$ which is manageable.

A larger problem with this method is that the Taylor-approximations above are approximations in the literal sense. We cannot generally say that the u following from the application of the FIC framework

generally converge towards the u_0 pursuant to traditional estimation unless we start out with asymptotically correct γ_0-values. Hence, the Taylor-approximation in this case are not approximations where the remainder in general can be claimed to converge to zero in probability, as is crucial when using Taylor developments to prove the asymptotic normality of the MLE, or in the derivation of FIC. The reason for this is the particular model assumptions imposed by the FIC-framework. This issue will be discussed further and illustrated in the simulation experiment in Chapter 6.

In our situation we are, however, not interested in the expected loss associated with a particular covariate combination, but the average loss over each covariate in the sample. Hence, if $x_1, ...x_n$, constitute the covariate combination in the sample, we are interested in estimating

$$E[L(\hat{\mu}_{n,M_i} - \mu_{true})] = \frac{1}{n}\sum_{j=1}^{n} E[L(\hat{\mu}_{n,M_i}(x_j) - \mu_{true}(x_j))]$$

Using the the second order Taylor development in equation (5.10) , we get

$$\begin{aligned} E[L(\hat{\mu}_{n,M_i} - \mu_{true})] \approx \frac{1}{n}\sum_{j=1}^{n} \{&L(u_{0i}) - u_{0i}L'(u_{0i}) + \frac{1}{2}u_{0i}^2 L''(u_{0i}) + (L'(u_{0i}) - u_{0i}L''(u_{0i}))BIAS(\hat{\mu}_{n,M_i}(x_j)) \\ &+ \frac{1}{2}L''(u_{0i})MSE(\hat{\mu}_{n,M_i}(x_j))\} \end{aligned}$$

Where u_{0i} is the initial value used for Taylor development. The corresponding estimates becomes

$$\begin{aligned} \hat{E}_n[L(\hat{\mu}_{M_i} - \mu_{true})] = \frac{1}{n}\sum_{j=1}^{n} \{&L(u_{0i}) - u_{0i}L'(u_{0i}) + \frac{1}{2}u_{0i}^2 L''(u_{0i}) + (L'(u_{0i}) \\ &- u_{0i}L''(u_{0i}))\widehat{BIAS}_n(\hat{\mu}_{n,M_i}(x_j)) + \frac{1}{2}L''(u_{0i})\widehat{MSE}_n(\hat{\mu}_{n,M_i}(x_j))\} \end{aligned}$$

We can use the FIC framework to to find expressions for $\widehat{BIAS}_n(\hat{\mu}_{M_i}(x_j))$ and $\widehat{MSE}_n(\hat{\mu}_{M_i}(x_j))$. Note, however, that in deriving the FIC we use an estimate for $\widehat{BIAS}^2(\hat{\mu}_{M_i}(x_j))$, which require an estimate of $\delta\delta^t$. For that we used $\hat{\delta}_{n,wide}\hat{\delta}_{wide}^t - \hat{Q}$ as explained in section 3.3.2. When estimating only δ, we can use $\hat{\delta}_{n,wide}$, although this is not fully satisfactory. This gives us

$$\widehat{BIAS}_n(\hat{\mu}_{n,M_i}(x_0)) = \frac{1}{\sqrt{n}}(\hat{\omega}_n(x_0)^t(\hat{\delta}_{n,wide} - \widehat{G}_{n,M_i}\hat{\delta}_{n,wide}))$$

We see that unless we use the same u_{0i} for all models we cannot use AFIC directly as an alternative to averaging over the FIC values. Hence, under this approach we cannot generally exploit the benefits of AFIC.

Remark 5.1. As we saw section 5.1.3, the loss function of the type $L(\mu, \hat{\mu}_{n,M_i}) = L(\hat{\mu}_{n,M_i} - \mu_{true})$ is quite general and applies to many loss functions, which makes this method rather generally applicable. It is for instance applicable for the useful LINEX loss function. However, as we will see below, for the LINEX, we have a more direct method available, which is better taking into account the cautions that must be taken with the Taylor-development method. Note, also, that there are practical loss functions outside this framework such as the Stein's loss function described in section 5.1.3. Another example is loss functions of the type

$$L(\mu_{true}, \hat{\mu}_{n,M_i}) = L(e^{\hat{\mu}_{n,M_i}} - e^{\mu_{true}})$$

for instance $L(\mu_{true}, \hat{\mu}_{n,M_i}) = (e^{\hat{\mu}_{n,M_i}} - e^{\mu_{true}})^2$. The framework here cannot be applied directly to such loss functions.

5.5.3 Direct use of FIC in the case of LINEX loss

In the case of LINEX loss, the FIC framework allows us to find a better approximation to the expected loss than using Taylor developments and in some cases even the exact expected loss. We will first present the principles, including taking some shortcuts, before we will return to some complications and observations in remarks. See also Hjort and Claeskens (2008) for a brief but similar discussion on the extention of the FIC-framework to the LINEX loss function.

For the use of LINEX we need some simple results. Let $X \sim N(\mu, \sigma^2)$ and $L(X) = c + b(e^{aX} - aX - 1)$, i.e the LINEX loss associated with X. Then by straightforward calculations for log-normal distributions, we have

$$E[L(X)] = c + b(e^{a\mu + \frac{1}{2}a^2\sigma^2} - a\mu - 1)$$

From the FIC framework we know that

$$\sqrt{n}(\hat{\mu}_{n,M_i} - \mu_{true}) \xrightarrow{d} \Lambda_{M_i} = \Lambda_0 + \omega^t(\delta - G_{M_i}D) \tag{5.11}$$

Where $\Lambda_{M_i}$is normal.

By slightly abusing notation, we have that

$$\sqrt{n}(\hat{\mu}_{n,M_i} - \mu_{true}) \xrightarrow{d} N(\sqrt{n}BIAS(\hat{\mu}_{n,M_i}), nVAR(\hat{\mu}_{M_i}))$$

where

$$VAR(\hat{\mu}_{n,M_i}) = \frac{1}{n}(\tau_0^2 + \omega^t Q_{M_i}^0 \omega)$$

$$\begin{aligned} BIAS(\hat{\mu}_{n,M_i}) &= E[(E(\hat{\mu}_{n,M_i}) - \mu_{true})] \\ &= \frac{1}{\sqrt{n}}(\omega^t(\delta - G_{M_i}\delta)) \end{aligned}$$

Hence, we have in the limit

$$(\hat{\mu}_{n,M_i} - \mu_{true}) \sim N(BIAS(\hat{\mu}_{n,M_i}), VAR(\hat{\mu}_{M_i}))$$

The LINEX loss is

$$L(\hat{\mu}_{n,M_i} - \mu_{true}) = c + b(e^{a(\hat{\mu}_{n,M_i} - \mu_{true})} - a(\hat{\mu}_{n,M_i} - \mu_{true}) - 1)$$

Consequently, we have under adequate assumptions in the limit that

$$E[L(\hat{\mu}_{n,M_i} - \mu_{true})] = c + b(e^{aBIAS(\hat{\mu}_{n,M_i}) + \frac{1}{2}a^2 VAR(\hat{\mu}_{n,M_i})} - aBIAS(\hat{\mu}_{n,M_i}) - 1) \tag{5.12}$$

The estimate of the expected loss, taking the average over covariates, becomes

$$\begin{aligned} \hat{E}_n[L(\hat{\mu}_{n,M_i} - \mu_{true})] &= \frac{1}{n}\sum_{j=1}^{n} \{c + b(e^{a\widehat{BIAS}_n(\hat{\mu}_{n,M_i}(x_j)) + \frac{1}{2}a^2\widehat{VAR}_n(\hat{\mu}_{n,M_i}(x_j))} \\ &\quad - a\widehat{BIAS}_n(\hat{\mu}_{n,M_i}(x_j)) - 1)\} \end{aligned} \tag{5.13}$$

Hence, we can use the results from the FIC framework to find an estimate of the expected loss.

Remark 5.2. The steps above may appear rather quick, and they indeed are. Note that in general we have that if $X_n \xrightarrow{d} X$ then $E(h(X_n)) \to E(h(X))$ for h continuous and bounded.[65] This is however to strict assumptions to be applied directly in our case as the LINEX loss function, $L(X)$, is continuous, but not bounded. Hence, we had to restrict equation (5.12) to adequate assumptions. As noted in Polansky (2011) p. 176 boundedness of h is not a necessary condition for the result to hold. We will not delve into the details here but note that as long as equation (5.11) behaves reasonably well we are on safe grounds. We will, however, avoid this problem, by instead rely on the limit distribution (as we do in FIC) for the limit results. By the continuous mapping theorem, we have that if $X_n \xrightarrow{d} X$ then $h(X_n) \xrightarrow{d} h(X)$ for h continuous.[66] Hence, we have that

$$L(\sqrt{n}(\hat{\mu}_{n,M_i} - \mu_{true})) \xrightarrow{d} L(\Lambda_{M_i}) = L(\Lambda_0 + \omega^t(\delta - G_{M_i}D)) \tag{5.14}$$

which leads us to the desired limit results based on the limit distribution.

Remark 5.3. The previous remark showed us that FIC an LINEX appears to works theoretically very well together. This provides an additional value of FIC, because LINEX is a flexible, adaptible and useful utility function. Hence the benefits of FIC can be extended to a range of situations where the loss of using a wrong model is different from the squared error. It will be beyond the scope of this study to go

[65] See Polansky (2011) Theorem 4.6 p. 172 for a precise formulation of the theorem and a proof.
[66] See, for instance, Knight (2000) p. 130.

further into the theoretical details in combining FIC and LINEX. We will however suggest what could be done.

An interesting study would be to create very simple examples to illustrate the combination of LINEX and FIC, such as the one in Claeskens and Hjort (2008) section 5.3 on "a precise tolerance limit" to analyze how the framework works. Basically Claeskens and Hjort (2008) assume two alternative models: the narrow and the wide. Hence, we have a model on the form

$$f_n(y) = f(y, \theta_0, \gamma_0 + \frac{\delta}{\sqrt{n}})$$

Now, both parameters are scalars. Hence, there is one protected parameter, θ_0, and one free parameter, $\gamma_0 + \frac{\delta}{\sqrt{n}}$. By doing so we can get analytical expressions for the condition for when a narrow model model is better in minimizing the expected loss than the wide model, when the free parameter is (wrongly) set to γ_0. Hence, we will see exactly when a narrow wrong model outperforms a correct wide model (for a given n), because of the increased variance of the wide model. This can of course be extended to more dimensions. Then we will have tolerance regions instead of a tolerance limit, see Claeskens and Hjort (2008) section 5.4.

Furthermore, we see that equation (5.14) can not only be used to get an estimate of the expected loss associated with a wrong model, but to make inferences from the entire distribution of the loss function. Since we have the the limit distribution of the loss as given in equation (5.14), we can perform various tests and construct power functions as described in Claeskens and Hjort (2008) section 5.3 remark 5.2.

Remark 5.4. Finally, a remark on estimation. Basically, for a given covariate combination x_j, we have used that the estimated expected loss is

$$\hat{E}_n[L(\hat{\mu}_{n,M_i} - \mu_{true})] = c + b(e^{a\widehat{BIAS}(\hat{\mu}_{n,M_i}(x_j)) + \frac{1}{2}a^2\widehat{VAR}(\hat{\mu}_{n,M_i}(x_j))} - a\widehat{BIAS}(\hat{\mu}_{n,M_i}(x_j)) - 1) \tag{5.15}$$

We know something about the properties of $\widehat{BIAS}(\hat{\mu}_{n,M_i}(x_j))$ and $\widehat{VAR}(\hat{\mu}_{n,M_i}(x_j)$. We know that that $nMSE(\hat{\mu}_{n,M_i}(x_j)) = n\widehat{BIAS}(\hat{\mu}_{n,M_i}(x_j))^2 + n\widehat{VAR}(\hat{\mu}_{n,M_i}(x_j)$ is asymptotically unbiased when $\hat{\delta}_{n,wide}\hat{\delta}^t_{n,wide} - \hat{Q}$, is used as an estimator for $\delta\delta^t$ as presented in section 3.3.2.[67] However, things becomes more problematic for $\sqrt{n}\widehat{BIAS}(\hat{\mu}_{n,M_i}(x_j))$ as we have no consistent estimate for δ. Generally, we don't know if the estimate in equation (5.15) is biased, and how large this potential bias is. We don't even know if it is consistent. This is an interesting issue for further research.

5.5.4 The use of FIC in prediction settings

As mentioned above FIC and AFIC are appropriate when we want to estimate the expected loss of a model and find the information-value of covariates when the focus is an estimate. However, by using

[67] See Hjort and Claeskens (2008) p. 150.

certain statistical results, we see that FIC and AFIC can be employed to estimate the expected loss also in certain prediction settings. Recall that from equation (3.11) that when we can write the model on the form $Y = g(x) + \varepsilon$,where the $\varepsilon's$ are iid $(0, \sigma^2)$, and with an estimated regression function $\hat{y}_{n,M_i,new}(x_0) = \widehat{g}_{n,M_i}(x_0)$, we have the following prediction error

$$\begin{aligned} MSE(\hat{y}_{n,M_i,new}(x_0)) &= E[(\widehat{g}_{n,M_i}(x_0) - Y_{new})^2] \\ &= \sigma^2 + VAR(\widehat{g}_{n,M_i}(x_0)) + BIAS^2(\widehat{g}_{n,M_i}(x_0)) \end{aligned} \tag{5.16}$$

Now, let $g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = x^t\beta$. We see that this model fits the FIC framework as a normal standard regression model. We can now let $\mu_{true} = E(Y_{new} \mid x) = g(x)$ and $\hat{\mu}_{n,M_i}(x) = \widehat{g}_{n,M_i}(x)$. We also see that the last two terms in equation (5.16) correspond to $MSE(\hat{\mu}_{n,M_i}(x))$. An estimate for $MSE(\hat{y}_{n,M_i,new}(x_0))$ is then

$$\widehat{MSE}_n(\hat{y}_{n,M_i,new}(x_0)) = \hat{\sigma}_n^2 + \widehat{MSE}_n(\hat{\mu}_{M_i}(x_0))$$

For the latter term, we can use the results from the FIC analysis. The challenge is to estimate σ^2. If we are only interested in calculating the information value of gathering covariates by measuring the difference in expected loss between models, we don't need to care about $\hat{\sigma}_n^2$. However, if we want to estimate $\hat{\sigma}_n^2$ there are reasonable estimates available. One alternative is to use the unbiased estimate obtained by using the full model, hence to let

$$\hat{\sigma}_n^2 = \hat{\sigma}_{n,wide}^2 = \frac{1}{n-(p+1)} \sum_{j=1}^{n} (\widehat{g}_{n,wide}(x_j) - y_j)^2$$

Since we are interested in the average MSE over the sample, we can either use the estimate

$$\widehat{MSE}1_n(\hat{y}_{n,M_i,new}) = \hat{\sigma}_{n,wide}^2 + \frac{1}{n} \sum_{j=1}^{n} \frac{FIC(\hat{\mu}_{n,M_i}(x_j))}{n}$$

or exploit the benefit of AFIC in the modified AFICM explained above, giving us

$$\widehat{MSE}2_n(\hat{Y}_{M_i}) = \hat{\sigma}_{wide}^2 + \frac{1}{n} AFICM(\hat{\mu}_{n,M_i})$$

5.6 Chapter summary

We have now considered three alternatives to estimate the expected loss associated with the use of a model for prediction or estimation of a focus: direct estimation, estimation by cross-validation, and a FIC-inspired approach. In prediction settings, the CV approach seems to be the most generally applicable, while the FIC approach is fairly generally applicable in estimation settings combined with Taylor

developments. FIC is also paricullary useful in the combination with the LINEX loss function. In this case we can use FIC without going via Taylor developments. In the special, but still very widely used case of normal linear regression, FIC can be used in prediction settings.

Estimating the expected loss associated with a model is in the context of this study is just a mean to have an appropriate measure to trade off the information value of a covariate against the costs of gathering that covariate, as explained in section 3.5. In the next chapter we will show this trade-off by a simulation experiment.

6 Loss estimation and cost information-value trade-off illustrated by a simulation experiment

6.1 The experiment setup

6.1.1 The data generating process

For simplicity we will assume the DGP to be a normal linear regression model of the type

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

where $\varepsilon_i \sim iid\, N(0, \sigma^2)$.

We will use the following numerical true values for the parameters

$$\begin{aligned} \beta_0 &= 10 \\ \beta_1 &= 10 \\ \beta_2 &= 1 \\ \beta_3 &= 1 \\ \sigma^2 &= 5^2 \end{aligned}$$

We will simulate sample sizes of n=100 and n=1000, to see how sample size affect our inferences. The covariates are assumed to be fixed and non-random, but we have generated the covariate values by independently drawing x_1 from uniform [-5,5], x_2 from uniform [-5,5], and x_3 from $N(0,1)$.

In addition we have generated two covariates not included in the model:

$$\begin{aligned} x_4 &= x_1 + N(0,1) \\ x_5 &= x_2 + N(0,1) \end{aligned}$$

To avoid any confucion, we assume that this relation is not known to the statistican. Hence, we will not delve into the issue of structural models/state-space models.

To obtain a system of nested models, we can construct the following wide model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$$

where the numerical values are as given above, and in addition:

$$\begin{aligned} \beta_4 &= 0 \\ \beta_5 &= 0 \end{aligned}$$

Since the values of these parameters are zero, including these additional covariates don't change the true nature of the DGP.

To not make things too complicated we will consider the constant term β_0 term as protected. This means that we in principle have five covariates to choose from, and in total $2^5 = 32$ models to consider, ranging from $M_{\{0\}}$, only consisting of the constant term, to $M_{\{012345\}}$, where all the covariates are included. However, we will not consider models including both x_1 and x_4 and/or x_2 and x_5 as they are considered as subsitute covariates. We will explain the purpose of this just below. The full model, $M_{\{012345\}}$, will be kept as the wide model, when the wide model is needed, but is not considered to be one of the candidate models to choose among. This leaves us with 18 candidate models in addition to the wide model

$$\begin{array}{ccccc} M_{\{0\}} & M_{\{01\}} & M_{\{02\}} & M_{\{03\}} & M_{\{04\}} \\ M_{\{05\}} & M_{\{012\}} & M_{\{013\}} & M_{\{015\}} & M_{\{023\}} \\ M_{\{024\}} & M_{\{034\}} & M_{\{035\}} & M_{\{045\}} & M_{\{0123\}} \\ M_{\{0135\}} & M_{\{0234\}} & M_{\{0345\}} & & \end{array}$$

From the above setup, we can see where we are going. We see that x_1, x_2 and x_3 are decreasingly important for the numerical value of Y. In addition, the low spread in the normal distribution associated with x_3 (relative to the spread of the responce), is likely to increase the variance of its associated parameter estimate and, hence, reduce its information-value. This can most easily bee seen using analogy to to a simple linear normal regression with one covariate. Let us say we have the model

$$Y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i$$

where ε_i are iid $N(0, \sigma^2)$. It is straightforward to find that the MLE and associated variance for α_1 are[68]

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

[68] This is a basic result to be found in all introductory regression literature. See, for instance, Wasserman (2003) p. 210 f.

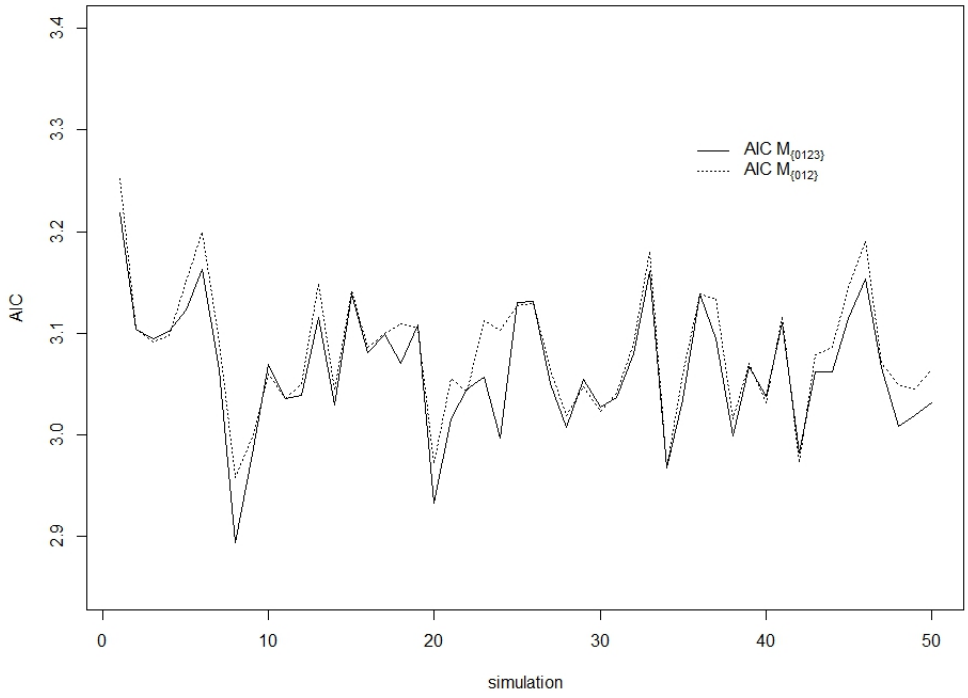

Figure 6.1: 100 simulation comparing AIC for $M_{\{012\}}$ and $M_{\{0123\}}$ for n=100

$$VAR(\hat{\alpha}_1) = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

We see that if the $x_i's$ are highly concentrated around $\bar{x}$, this increases the variance for $\hat{\alpha}_1$, for a given σ^2. Furthermore, we see that the variance reduces with n.

As a preliminary investigation, we have performed a simulation in Figure 6.1 showing that $AIC(M_{\{012\}})$ beats $AIC(M_{\{0123\}})$ relatively often when n=100 (AIC is defined to be on a loss scale and, hence, to be as low as possible).[69] From figure 6.1 we see that, although x_3 is included in the DGP, we often get simulated samples where it is not recommended to be included in the model according to AIC for n=100. Hence, as indicated above, the high variance associated with $\hat{\beta}_{3,n=100}$ is likely to give it low information-value. As we will see below, this changes as n increases to 1000.[70] Note that it might, on first sight,

[69] We have used that AIC=$\frac{-loglik+p}{2n}$, which is desired to be as low as possible.

[70] For a discussion on the relation between AIC an hypothesis testing, see Claeskens and Hjort (2008) p. 50.

appear rather artificial to present the different AIC values in a graph, as they are just different AIC values for replicated simulations with no natural order. We think however the graph is illustrative in illustrating winner models from the various simulations, and it also illustrates that AIC is indeed an estimate subject to variations. This will be commented upon in the concluding remarks.

From the above setup, we see that there is intuitively a strong correlation between x_1 and x_4, and x_2 and x_5. We can consider observing x_4 and x_5 as observing x_1 and x_2 with noise (although the structure of the noise is assumed unknown). The point here is that although x_4 and x_5 are not included in the model, they may be useful replacements for x_1 and/or x_2, if they are less costly to observe. The idea is that we can sacrify some preciseness if it reduces costs sufficiently.

Note that when applying the FIC framework in the experiment, we willl let $\gamma_0 + \delta/\sqrt{n} = (\beta_1, ..., \beta_5)^T$. For the purpose of this experiment we have set $\gamma_0 = 0$.

6.1.2 Foci to be considered

We will consider three foci. Our first focus is to predict a new Y, Y_{new}. The second focus is that we want to estimate the expectation of a new Y, $E(Y_{new})$. The third focus is the estimate the 5 percent percentile of a new Y, $F^{-1}_{Y_{new}}(0.05)$. Briefly, we can say that we are interested in 3 different "estimates" of a new Y: The predicted Y, the expected Y and the 5 percent percentile of Y, i.e.

$$\begin{aligned} \mu^1 &= Y_{new} \\ \mu^2 &= E(Y_{new}) \\ \mu^3 &= F^{-1}_{Y_{new}}(0.05) \end{aligned}$$

In line with the theoretical analysis above, we wish to estimate the expected loss for the foci averaged over all the covariate combinations in the sample, i.e.

$$\widehat{E}_n[L(\hat{\mu}^k_{n,M_i}, \mu^k) = \frac{1}{n}\sum_{j=1}^{n} \widehat{E}_n[L(\hat{\mu}^k_{n,M_i}(x_j), \mu^k(x_j))]$$

As just mentioned, we have chosen foci that are all related to the estimation of a new Y. By doing this all foci are on the same scale which makes the expected loss on the same scale for a given loss function. This is useful in comparing the information value of the covariates in terms of reduction in estimated expected loss.

For comparison, we have also computed the results for $F^{-1}_{Y_{new}}(0.01)$ and $F^{-1}_{Y_{new}}(0.95)$ that will be briefly referred to for comparison when relevant. The experimental results of for these foci are listed in Appendix B.

Remark 6.1. Note that for the FIC-framework, in the case of normal linear regression, ω is the same for $E(Y_{new} \mid x_j)$ and $F^{-1}_{Y_{new}}(\alpha \mid x_j)$. To see this note that

$$E(Y_{new} \mid x_j) = x_j^t \beta$$

and that

$$F^{-1}_{Y_{new}}(\alpha \mid x_j) = x_j^t \beta + \sigma \Phi^{-1}(\alpha) \tag{6.1}$$

where $\Phi^{-1}(\alpha)$ is the α-percentile of a $N(0,1)$. The $\frac{\partial \mu}{\partial \sigma}$ part drops out of the calculation of ω. See example 6.6 in Claeskens and Hjort (2008) p. 163. The only measurement that will separate the FIC for $E(Y_{new} \mid x_j)$ and $F^{-1}_{Y_{new}}(\alpha \mid x_j)$ for various α, is τ_0^2 , which is dependent on $\frac{\partial \mu}{\partial \sigma}$. Hence the FIC values will be equal except for a constant. Consequently, looking at FIC alone, one should believe that the information-value of a covariate should not depend on the focus in this context. This will be confirmed by our experiment when comparing μ^2 and μ^3.

Remark 6.2. Note that for the FIC framework, in the case of normal linear regression, the FIC based MSE for a focus linear in mean parameters, such that $E(Y_{new} \mid x_j)$, is exact, and not just an asymptotic approximation. A technical explanation for this is given in Claeskens and Hjort (2008) example 6.7. This is important, because we will then know that the results for such foci are not influenced by asymptotic approximation errors. This will in particular be useful when comparing the results for $\mu^2 = E(Y_{new})$ with $\mu^3 = F^{-1}_{Y_{new}}(0.05)$, because we then know that only the latter will be additionally disturbed by asymptotic approximation errors. However, this of course, does not mean that the estimated expected losses in our experiment are exact, since we deal with estimates, and not the exact measurements needed to find the bias and variance in the FIC framework.

Remark 6.3. Note that equation (6.1) combined with the fact that $\Phi^{-1}(\frac{1}{2}) = 0$, where $\Phi^{-1}(\alpha)$ is the α-percentile of a $N(0,1)$, reveals that as foci $E(Y_{new} \mid x_j)$ and $F^{-1}_{Y_{new}}(\frac{1}{2} \mid x_j)$ are identical.

Remark 6.4. Note that the error of using a wrong model for estimation may increase when we estimate extreme values, simply because the variance/standard deviation becomes more important for the estimation. This is easily seen by inspecting equation (6.1). When α is zero, the variance does not matter at all. We can also illustrate this by a simplified numerical example. Let us say that the true DGP for some data $N(0,1^2)$. This means that the true $E(Y) = 0$ and the true $F_Y^{-1}(0.05) = -1.65$. Let us assume that we use the model $N(0.01,2^2)$ to make parameter inferences. When using $N(0.01,2^2)$, we see that we wrongly estimate $E(Y)$ by 0.01. However $F_Y^{-1}(0.05)$ is wrongly estimated by 0.01-2*1.65+1.65, which is 1.64. Hence, the error of estimating $F_Y^{-1}(0.05)$ is obviously larger, even if the same model is used for estimation. Note that this example is not crucially dependent on that we estimate from a model with larger variance. Let us instead assume that we use the model $N(0.01,0.5^2)$ for estimation. Then the error in estimating $F_Y^{-1}(0.05)$ is 0.01-0.5*1.65+1.65=0.835.

In our setup we see that it is crucial to take into account the standard error of each model, to get a correct picture of the errors of the estimates. In table 6.1 we have shown how different the standard deviations becomes under the traditional estimation of all the models for n=1000. We see here that especially for those models missing the important covariate, x_1, or its substitute x_4, the standard deviation becomes huge, as the impact of these covariates are baked into the standard error. However, as we will see below, the FIC-framework will not capture this. In the FIC-framework the σ is estimated on the basis

	$\hat{\sigma}_{M_i,n=1000}$
$M_{\{0\}}$	29.6607
$M_{\{01\}}$	5.9376
$M_{\{02\}}$	29.6227
$M_{\{03\}}$	29.6541
$M_{\{04\}}$	11.4161
$M_{\{05\}}$	29.5871
$M_{\{012\}}$	5.2156
$M_{\{013\}}$	5.8394
$M_{\{015\}}$	5.3475
$M_{\{023\}}$	29.6160
$M_{\{024\}}$	11.0100
$M_{\{034\}}$	11.3714
$M_{\{035\}}$	29.5796
$M_{\{045\}}$	11.0791
$M_{\{0123\}}$	5.1002
$M_{\{0135\}}$	5.2275
$M_{\{0234\}}$	10.9621
$M_{\{0345\}}$	11.0279
$M_{\{012345\}}$	5.0957

Table 6.1: Estimated σ's under the various models

of the wide model, and used for all models, as it is asymptotically equivalent for all models under the FIC-framework.[71] The reason for this is lurking behind the assumption of the FIC model. First note that for normal linear regression the σ estimate is independent of the covariate parameter estimates under the true model, i.e, the wide model. To see the asymptotic equivalence, recall that the FIC-framework relies on the distribution function

$$f_n(y) = f(y \mid \theta_0, \gamma_0 + \delta/\sqrt{n})$$

As long as we operate within an environment where γ_0 are assigned its asymptotically correct value, then we can base the estimation of σ on any submodel.[72] However, recall that in the experiment, we have set $\gamma_0 = 0$, which is not asymptotically correct according to the experiment setup. In this setting, the use of the wide model for σ estimation in some sence misses the model dependency of σ. We will see that this will have a high impact on our experiment, as the FIC-framework will in some sence understate the real impact of the standard deviation on estimating the foci compared to traditional estimation methods (not via the FIC-framework).

[71] See Claeskens and Hjort (2008) p. 161.
[72] See Claeskens and Hjort (2008) p. 154.

6.1.3 Loss functions considered

For the prediction of a new $\mu^1 = Y_{new}$, we have chosen the squared error as the loss function. The reason is that this allows us to try all the methods for expected loss estimation explained in Chapter 5: we can use direct estimation by using plug-in estimates, the empirical distribution and bias correction. We can also use cross-validation, and we can use the adjusted FIC for prediction settings as described in section 5.5.4.[73] Hence, we will see if the methods used for loss estimation give approximately the same recommendations regarding the information value of covariates.

For the foci $\mu^2 = E(Y_{new})$ and $\mu^3 = F^{-1}_{Y_{new}}(0.05)$, we will use two parametric versions of the LINEX loss function. We will for simplicity use b=100 in both cases to get the measurements at an convenient scale. In one version we will use a=0.1, implying a strong asymmetric aversion towards positive errors, and in the other version, a=-0.1, implying a strong aversion for negative errors. Since we are now in an estimation setting, and not in a prediction setting, we will use FIC combined with Taylor development as explained in section 5.5.2, and use FIC estimates inserted in the expected loss as explained in section 5.5.3, which we will call the direct method, to estimate the expected loss and information value of covariates. Hence, we will be able to compare the results of the two approaches to estimate the expected loss associated with the models.

6.1.4 The information-value cost trade-off in the experiment

We will not set some fixed covariate costs, as fixing the costs will be arbitrarily, and does not provide us with any general results. Rather, we will focus on the information-value and compare between the models what costs are necessary to justify omitting informative covariates. We will mainly focus on two aspects regarding the inclusion of covariates.

The first aspect we will be concerned with is the value of including x_3. There are two a priori aspects that say that the information-value of this covariate is low. Firstly, x_3 has a low impact on Y as such. Secondly, as a consequence of our simulation setup, x_3 has a relatively low spread due to its distribution. As pointed out above, this tends to make the variance of $\hat{\beta}_3$ high compared to the uniform distribution used in simulating x_1 and x_2. In this case the number of observations must be increased to reduce the variance. For our simulated model this can bee seen from the parameter estimates estimated on the basis of the true model in Table 6.2. We see that for n=100, then $\hat{\beta}_3$ is not significant. However, it becomes significant when n=1000.

The second aspect we will be concerned with is how much more are we willing to pay for preciseness. Both x_1 and x_2, have the less precise alternatives x_4 and x_5, respectively. Hence, the question is how much the inclusion of the presice terms improves information value compared to the inclusion of the noisy terms. Since x_1 as such is more important in determining the size of Y than x_2, we would a priori believe that we are willing to pay more to have x_1 precise than x_2.

[73] Note that in the experiment we have used the truncated version of the estimated bias squared in the application of FIC as described in section 3.3.2.

	$\hat{\beta}_{i,n=100}$	$s.e.(\hat{\beta}_{i,n=100})$	$\frac{\hat{\beta}_{i,n=100}}{s.e.(\hat{\beta}_{i,n=100})}$	$\hat{\beta}_{i,n=1000}$	$s.e.(\hat{\beta}_{i,n=1000})$	$\frac{\hat{\beta}_{i,n=1000}}{s.e.(\hat{\beta}_{i,n=1000})}$
$\hat{\beta}_0$	9.7523	0.5687	17.1486	9.9490	0.1613	61.6613
$\hat{\beta}_1$	10.1146	0.2149	47.0657	10.0079	0.0553	180.8844
$\hat{\beta}_2$	0.7661	0.1953	3.9235	0.9997	0.0567	17.6321
$\hat{\beta}_3$	0.3295	0.5752	0.5728	1.0459	0.1546	6.7663

Table 6.2: Parameter estimates and s.e. under the true model

The inclusion of covariates will be studied along several lines. We will see how the conclusions are affected by the technique used for expected loss estimation. Furthermore, we will see how, the information-value depends on the focus chosen, and the impact of asymmetry of the loss function. Finally, we will see how the conclusions are affected by n, by comparing n=100 and n=1000. We will also see how increasing n affects the estimated expected loss of the winner model. An additional aspect that could have been analyzed is variations in the parameters in the LINEX loss function, especially variations in a. A preliminary investigation showed us that this would not change much in model ranking, but naturally rescale the information value of covariates. Our assessment is that extending the analysis to also include variations in a will not add much to the analysis as a is chosen at a suitable level to get the losses at a reasonable scale. Analyzing variations in a at this level would not provide much general insight. In a real world setting the statistician should, as a part of the sensitivity analysis, check how small variations in a is likely to affect the optimality of gathering certain covariates.

Note that in this experiment we will implicitly assume that we have to decide all the covariates to be gathered at once. In many applications this will be a very strict assumption, as one sequentially can gather covariates and then decide what additional covariates to gather. It is for instance easy to imagine that if we first have gathered x_1, then the information value of gathering the other covariates is altered depndent on the value of x_1. Hence, we need some sort of sequential optimization. We will briefly return to some aspects of sequential covariate gathering in the discussion of μ^3.

6.2 Experiment results and analysis for μ^1

6.2.1 The result of the experiment

We have estimated the expected mean square prediction error of the candidate models using four methods for n=100 and n=1000 in Table 6.3 and Table 6.4, respectively. The empirical estimate is the bias corrected empirical estimate derived in section 5.3.2. The CV estimate is the leave-one-out cross-validation estimate derived in section 5.4.1. The average FIC utilizes the FIC framework for prediction as described in section 5.5.4 where we average over the FIC results for each covariate. The AFICM also uses the FIC framework, but utilizes the modified AFIC, AFICM, as described in section 5.5.1.

Note that many probably find it more appealing to express such tables on $\sqrt{MSE}$ scale rather than the

	$\widehat{MSE}^{emprical}_{n=100}$		$\widehat{MSE}^{CV}_{n=100}$		$\widehat{MSE}^{averageFIC}_{n=100}$		$\widehat{MSE}^{AFICM}_{n=100}$	
$M_{\{0\}}$	716.8296	17	730.7713	13	717.0402	17	716.9446	17
$M_{\{01\}}$	36.7877	5	36.8772	5	37.1130	5	36.9027	5
$M_{\{02\}}$	713.7334	14	742.5660	15	713.8728	14	713.8484	14
$M_{\{03\}}$	715.6511	15	742.4946	14	715.8113	15	715.7661	15
$M_{\{04\}}$	109.7188	12	112.6349	12	109.8619	12	109.8338	12
$M_{\{05\}}$	717.0705	18	745.9703	16	717.2684	18	717.1855	18
$M_{\{012\}}$	32.6094	1	32.6151	1	33.1063	1	32.8498	1
$M_{\{013\}}$	37.2356	6	37.4809	6	37.4787	6	37.3506	6
$M_{\{015\}}$	33.0123	2	33.1828	3	33.4445	3	33.1273	2
$M_{\{023\}}$	712.7198	13	754.5644	17	712.8972	13	712.8348	13
$M_{\{024\}}$	106.7384	9	111.4125	9	106.9125	9	106.8534	9
$M_{\{034\}}$	107.5101	11	111.8017	10	107.6401	11	107.6251	11
$M_{\{035\}}$	715.9562	16	758.0254	18	716.1017	16	716.0712	16
$M_{\{045\}}$	107.3775	10	111.9941	11	107.5454	10	107.4925	10
$M_{\{0123\}}$	33.1093	3	33.1234	2	33.4005	2	33.2243	3
$M_{\{0135\}}$	33.5242	4	33.7156	4	33.7883	4	33.6392	4
$M_{\{0234\}}$	104.7342	7	110.6750	7	104.8598	7	104.8492	7
$M_{\{0345\}}$	105.3999	8	111.2884	8	105.5331	8	105.5149	8

Table 6.3: Estimated mean square prediction error n=100 with ranking of models

MSE scale, to get the measurements on the same scale as the data. This will not alter model selection based on model ranking. This will, however, change the information-value of the covariates. For the purpose of this study, MSE makes more sense as a loss function than taking the root, as MSE reflects risk aversion as described in Chapter 5.

6.2.2 Comparing the estimation methods

We see that the four different methods yield approximately the same estimated expected losses. This indicates that all methods are applicable to estimate the expected loss associated with the various models and calculate the information value of covariates.

The CV estimates seem to be most different from the other methods, especially for those models most deviant to the true DGP, but the estimates from the various methods becomes more aligned as n increases from 100 to 1000. As described in Section 5.4, an explanation for this is that the CV estimate is slightly different to the other methods. Firstly, there is a learning curve effect that might bias the estimate upwards, because we use one less observation in the parameter estimation. Whe cannot exclude that this has a noticeable effect, at least for n=100. Furthermore, the CV takes into account out-of-sample bias with respect to the covariates, while in the other methods the covariates are held fixed and assumed

	$\widehat{MSE}^{emprical}_{n=1000}$		$\widehat{MSE}^{CV}_{n=1000}$		$\widehat{MSE}^{averageFIC}_{n=1000}$		$\widehat{MSE}^{AFICM}_{n=1000}$	
$M_{\{0\}}$	879.8110	18	881.5212	16	879.8125	18	879.8119	18
$M_{\{01\}}$	35.3590	6	35.3944	6	35.3635	6	35.3600	6
$M_{\{02\}}$	877.6103	16	881.1023	15	877.6115	16	877.6112	16
$M_{\{03\}}$	879.4714	17	882.8708	18	879.4724	17	879.4723	17
$M_{\{04\}}$	130.4302	12	130.8350	12	130.4320	12	130.4311	12
$M_{\{05\}}$	875.5029	14	878.9388	13	875.5043	14	875.5038	14
$M_{\{012\}}$	27.3588	2	27.3640	2	27.3659	2	27.3597	2
$M_{\{013\}}$	34.2548	5	34.2983	5	34.2580	5	34.2557	5
$M_{\{015\}}$	28.7520	4	28.7648	4	28.7584	4	28.7529	4
$M_{\{023\}}$	877.2613	15	882.4495	17	877.2625	15	877.2623	15
$M_{\{024\}}$	121.3754	8	121.9586	8	121.3770	8	121.3763	8
$M_{\{034\}}$	129.4649	11	130.0766	11	129.4664	11	129.4659	11
$M_{\{035\}}$	875.1102	13	880.2386	14	875.1114	13	875.1111	13
$M_{\{045\}}$	122.9022	10	123.4899	10	122.9043	10	122.9032	10
$M_{\{0123\}}$	26.2198	1	26.2150	1	26.2391	1	26.2267	1
$M_{\{0135\}}$	27.5342	3	27.5409	3	27.5375	3	27.5351	3
$M_{\{0234\}}$	120.3751	7	121.1484	7	120.3762	7	120.3760	7
$M_{\{0345\}}$	121.8226	9	122.6037	9	121.8240	9	121.8236	9

Table 6.4: Estimated mean square prediction error n=1000 with ranking of models

non-random. This is likely to increase the estimated expected loss. The effect of this is diminishing as n increases as the two measurements converge.

For n=100 all methods picks the same winner model $M_{\{012\}}$, but very small differences in the estimates cause some disagreement on the choice of model 2. Except for this, the methods agree on the ranking up to the 10th best model. When n=1000 all methods agree on the true DGP, $M_{\{0123\}}$, as the winner model, and the ranking up to the 13th best model.

6.2.3 The information value of x_3

When n=100, the model $M_{\{012\}}$ is ranked as number one. The estimated expected loss increases by including x_3. This is not surprising since $\hat{\beta}_3$ is not significant. Including it increases variance more that it reduces squared bias. Hence, even if it was free to gather x_3, it should not be included. As pointed out in Section 3.5, Remark 3.1, it is not always rational to take information into account, even if it is free. This changes when n=1000. $M_{\{0123\}}$ is now picked out as the winner model, with $M_{\{012\}}$ as the second. Recall that from above that $\hat{\beta}_3$ is now significant and is estimated with a lower variance. The estimated mean squared error seems to be approaching the variance $\sigma^2 = 25$, which is the irreducible error of prediction.

We see that for all estimation methods adding x_3 to $M_{\{012\}}$ reduce the expected loss by little more

than one for n=1000. Hence, if, for instance, the cost c_3 of gathering x_3 is two, it should not be gathered. Note that it if c_3 is sufficient high, it will never be optimal to gather x_3, even for very large n. The reason is that the reduced bias by including x_3 is never worth the cost. Hence, when incorporating costs, it may never be optimal to use the true DGP!

6.2.4 The value of more precise information

We can now look at the value of more precise information. Let us first look at how much we gain from gathering x_1 instead of the cheaper substitute x_4. We see that both for n=100 and n=1000, the increased expected loss of substituting x_1 by x_4 is substantial. If we, when n=100, substitute $M_{\{012\}}$ by $M_{\{024\}}$, we see that the estimated expected loss increases from 32.61 to 106.7, using the empirical distribution bias adjusted estimate (the number will be almost the same using the other estimates). If we, when n=1000, substitute $M_{\{0123\}}$ by $M_{\{0234\}}$ the estimated expected loss increases from 26.22 to 120.34 using the empirical distribution bias adjusted estimate. Hence the information value of precision of x_1 is high. This is not surprising since x_1 is so important for the measurement of Y. It is also not surprising that the value becomes even higher as n increases, as we then get more precise estimates.

Substituting x_2 by x_5, doesn't have the same impact. This is not surprising as x_2 is less important for the nominal value of Y than x_1. If we, when n=100, substitute $M_{\{012\}}$ by $M_{\{015\}}$, we see that the estimated expected loss increases from 32.61 to 33.01 using the empirical bias adjusted estimate. If we, when n=1000, substitute $M_{\{0123\}}$ by $M_{\{0135\}}$ the estimated expected loss increases from 26.22 to 27.53, using the empirical distribution bias adjusted estimate.

6.2.5 The value of increased amount of data

We see that increasing n form 100 to 1000 has quite large effect on the estimated expected loss. If we use the direct estimates an example we see that for the winner model, $M_{\{012\}}$, the estimated expected loss is 32.61, while the estimated expected loss is 26.22 for the winner model, $M_{\{0123\}}$, when n=1000. As n increases, and assuming the model with lowest estimated expected loss is used, the estimated expected loss approaches 25, which is the theoretical minimum following from irreducible error. Taking into account the limit properties of the estimators, this is as expected.

6.3 Experiment results for μ^2

6.3.1 The result of the experiment

The results for $E(Y_{new})$ for n=100 and n=1000 are given in Table 6.5 and Table 6.6, respectively. The Taylor-based estimate is the estimate developed from the second order Taylor-development of the loss function, and using FIC, explored in Section 5.5.2, while the direct estimate is based on the direct method applicable to LINEX loss functions explored in Section 5.5.3.

	$\widehat{EL}^{Taylor}_{n=100,a=0.1}$		$\widehat{EL}^{Direct}_{n=100,a=0.1}$		$\widehat{EL}^{Taylor}_{n=100,a=-0.1}$		$\widehat{EL}^{Direct}_{n=100,a=-0.1}$	
$M_{\{0\}}$	870.3478	16	880.0381	16	1236.2364	16	1247.7644	17
$M_{\{01\}}$	2.5765	5	3.0984	5	2.6195	5	3.1296	5
$M_{\{02\}}$	851.7868	13	859.7528	13	1237.9751	17	1247.2135	16
$M_{\{03\}}$	889.1110	18	896.7961	18	1220.7620	13	1231.1844	13
$M_{\{04\}}$	45.4529	11	46.4327	12	52.0179	12	53.2013	12
$M_{\{05\}}$	867.0019	14	875.0023	14	1240.0995	18	1249.6297	18
$M_{\{012\}}$	0.5757	1	0.8448	1	0.5814	1	0.8394	1
$M_{\{013\}}$	2.7778	6	3.1860	6	2.7986	6	3.1970	6
$M_{\{015\}}$	0.7445	3	1.0510	3	0.7517	3	1.0389	3
$M_{\{023\}}$	870.1390	15	876.1873	15	1231.5331	15	1239.8521	15
$M_{\{024\}}$	45.3677	10	46.1245	10	46.6490	8	47.4851	8
$M_{\{034\}}$	43.3265	8	44.0805	8	51.6368	11	52.6056	11
$M_{\{035\}}$	886.0181	17	892.1251	17	1227.3907	14	1235.9595	14
$M_{\{045\}}$	45.5581	12	46.3159	11	47.8828	10	48.7458	10
$M_{\{0123\}}$	0.7261	2	0.9459	2	0.7271	2	0.9399	2
$M_{\{0135\}}$	0.9199	4	1.1568	4	0.9229	4	1.1471	4
$M_{\{0234\}}$	43.2094	7	43.7649	7	46.0371	7	46.6777	7
$M_{\{0345\}}$	43.5482	9	44.0958	9	47.2249	9	47.8923	9

Table 6.5: Estimated LINEX loss for E(Y), b=100, a=0.1 and a=-0.1 for n=100

6.3.2 Comparing the estimation methods

We see that the two methods, Taylor-based estimation and direct estimation seems to yield fairly similar results, both for $a = 0.1$ and $a = -0.1$. However, it seems like the Taylor-based method systematically gives somewhat lower estimates than the direct method. The impact of this seems to be smaller as n increases for 100 to 1000, though, indicating this is less of a problem as n grows.

As the estimates becomes more precise as n increases, the difference in estimated expected loss between the good models and the poorer models increases. We have no reasons to believe that the average estimated expected losses should be influenced by the asymmetry in the loss functions, i.e whether $a = 0.1$ or $a = -0.1$. This is supported by the data. This will be discussed further when analyzing μ^3, and there we will see that the sign of a will influence the estimated expected loss when analyzing a single covariate combination.

We see that the estimated expected loss boost when omitting the central covariate x_1, or its substitute x_4. This is not surprising. As this is the most important covariate in providing information about Y, removing is likely to yield very imprecise models. We will elaborate further on this topic when analyzing μ^3.

	$\widehat{EL}^{Taylor}_{n=1000,a=0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=0.1}$		$\widehat{EL}^{Taylor}_{n=1000,a=-0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=-0.1}$	
$M_{\{0\}}$	1478.6595	18	1479.9265	18	1483.3746	18	1484.6319	18
$M_{\{01\}}$	4.7021	6	4.7558	6	4.6881	6	4.7409	6
$M_{\{02\}}$	1447.9046	16	1448.9427	16	1446.0152	16	1447.0237	16
$M_{\{03\}}$	1471.8955	17	1472.9577	17	1471.4647	17	1472.5098	17
$M_{\{04\}}$	71.5231	12	71.6372	12	68.0454	12	68.1552	12
$M_{\{05\}}$	1447.1380	15	1448.1838	15	1431.8338	14	1432.8382	14
$M_{\{012\}}$	0.6269	2	0.6635	2	0.6196	2	0.6556	2
$M_{\{013\}}$	4.1120	5	4.1519	5	4.1228	5	4.1626	5
$M_{\{015\}}$	1.3404	4	1.3780	4	1.3130	4	1.3496	4
$M_{\{023\}}$	1440.4946	14	1441.3319	14	1434.4340	15	1435.2328	15
$M_{\{024\}}$	62.2514	8	62.3350	8	59.8317	8	59.9122	8
$M_{\{034\}}$	70.3091	11	70.3997	11	67.3044	11	67.3920	11
$M_{\{035\}}$	1439.3423	13	1440.1874	13	1419.5492	13	1420.3433	13
$M_{\{045\}}$	63.5577	10	63.6434	10	61.4574	10	61.5378	10
$M_{\{0123\}}$	0.0582	1	0.0750	1	0.0581	1	0.0749	1
$M_{\{0135\}}$	0.7132	3	0.7385	3	0.7063	3	0.7313	3
$M_{\{0234\}}$	61.0549	7	61.1165	7	59.1311	7	59.1906	7
$M_{\{0345\}}$	62.3247	9	62.3880	9	60.6111	9	60.6703	9

Table 6.6: Estimated LINEX loss for E(Y), b=100, a=0.1 and a=-0.1 for n=1000

6.3.3 The information value of x_3

As for the analysis of the Y_{new} prediction, the model $M_{\{012\}}$ is ranked as number one for n=100. The estimated expected loss increases by including x_3, and hence x_3 should not be gathered no matter the cost. When n=1000 the estimated expected loss deceases by including x_3.

The information value of including x_3 seems to not be to a noticeable extent be affected by the method used for estimation or the asymmetry of the loss function. We see that the value of x_3 is around 0.57, regardless of estimation method and asymmetry. If we use the direct method and $a = 0.1$ as an example, we see that adding x_3 to $M_{\{012\}}$, reduces the estimated expected loss from 0.66 to to the estimated loss 0.08 of the winner model $M_{\{0123\}}$. When using the Taylor method the same estimated estimated expected loss reduces from 0.63 to 0.06. It might also instructive to observe that omitting x_3 multiplies the expected loss by little less than 10, when direct estimation is used, and little more that 10 when using Taylor based estimates.

For c_3 sufficient large, let us say 1, it is not cost-efficient to gather x_3 when n=1000, and hence, probably not for any n.

In the discussion of μ^3, we will see that the the information value of x_3 may be altered if we take into account the possibility of sequential covariate gathering.

6.3.4 The value of more precise information

We see that it is hardly any alternative to gather x_4 as a substitute for x_1 as long as not c_1 is extremely high (relative to c_4). However, gathering x_5 instead of x_2 seems to be a more practical alternative, especially when n=100. In the case of n=100, gathering x_5 instead of x_2 seems to only increase the estimated expected loss by 0.20 when using $M_{\{024\}}$ instead of the winner model $M_{\{012\}}$ regardless of estimation method and asymmetry. If we, when n=1000, substitute $M_{\{0123\}}$ by $M_{\{0135\}}$ the estimated expected loss increases by approximately 0.65. This reflects that when n=100 there are more uncertainty associated with estimating the regression parameter for x_2 in the first place, and we don't do so much worse by using the replacement x_5 instead. Hence, the value of more precise information increases as n increases.

6.3.5 The value of increased amount of data

We see that increasing n from 100 to 1000 has quite a substantial effect on the estimated expected loss. This effect is not crucially dependent on the method used or the value of a. If we use the direct estimate for $a = 0.1$ as an example we see that for the winner model, $M_{\{012\}}$, the estimated expected loss is 0.58, while the estimated expected loss is 0.06 for the winner model, $M_{\{0123\}}$, when n=1000.

6.4 Experiment results for μ^3

6.4.1 The result of the experiment

The results for $F^{-1}_{Y_{new}}(0.05)$ for n=100 and n=1000 are given in Table 6.7 and Table 6.8, respectively. The Taylor-based estimate is the estimate developed from the second order Taylor-development of the loss function, and using FIC, explored in Section 5.5.2, while the direct estimate is based on the direct method applicable to LINEX loss functions explored in Section 5.5.3.

6.4.2 Comparing the estimation methods

At an overall level we see that this focus doesn't significantly alter model selection compared to the foci considered above.With respect to $\mu^2 = E(Y_{new})$, this is no surprise at all, taking into account Remark 6.1. However, the Taylor approximation method seems to produce rather extreme results for the poor models. Things seems to really start to go wrong for models omitting the important covariate x_1, or its substitute, x_4. This is not surprising, as explained above, as x_1 is the most important covariate in providing information about Y, removing it is likely to yield very imprecise models.

However, to explain the extreme results specifically when using the second order Taylor approximation estimate, we must go deeper into the Taylor series and the model. For positive a the Taylor method seems to underestimate the expected loss, while for negative a the Taylor method seems to overestimate the expected loss. To explain this, first recall that a Taylor development of the loss function using the

	$\widehat{EL}^{Taylor}_{n=100,a=0.1}$		$\widehat{EL}^{Direct}_{n=100,a=0.1}$		$\widehat{EL}^{Taylor}_{n=100,a=-0.1}$		$\widehat{EL}^{Direct}_{n=100,a=-0.1}$	
$M_{\{0\}}$	213.6687	15	882.0309	16	161395.0604	18	1250.5049	17
$M_{\{01\}}$	2.7775	5	3.3080	5	2.8384	5	3.3393	5
$M_{\{02\}}$	208.8075	13	861.7043	13	158149.7995	16	1249.9529	16
$M_{\{03\}}$	219.4427	18	898.8229	18	157898.7566	14	1233.8912	13
$M_{\{04\}}$	38.7231	10	46.7304	12	77.1619	12	53.5128	12
$M_{\{05\}}$	211.8438	14	876.9848	14	161216.4380	17	1252.3740	18
$M_{\{012\}}$	0.7752	1	1.0498	1	0.7900	1	1.0445	1
$M_{\{013\}}$	2.9592	6	3.3959	6	3.0380	6	3.4068	6
$M_{\{015\}}$	0.9427	3	1.2565	3	0.9627	3	1.2444	3
$M_{\{023\}}$	214.4334	16	878.1722	15	155930.1026	13	1242.5765	15
$M_{\{024\}}$	39.0134	11	46.4216	10	68.5298	9	47.7850	8
$M_{\{034\}}$	36.9426	7	44.3735	8	74.7757	11	52.9159	11
$M_{\{035\}}$	217.7204	17	894.1424	17	158127.3556	15	1238.6760	14
$M_{\{045\}}$	39.0872	12	46.6134	11	70.4394	10	49.0483	10
$M_{\{0123\}}$	0.9240	2	1.1512	2	0.9369	2	1.1451	2
$M_{\{0135\}}$	1.1149	4	1.3625	4	1.1367	4	1.3528	4
$M_{\{0234\}}$	37.1706	8	44.0573	7	66.2492	7	46.9760	7
$M_{\{0345\}}$	37.3796	9	44.3888	9	68.0754	8	48.1931	9

Table 6.7: Estimated LINEX loss for $F^{-1}(0.05)$, b=100, a=0.1 and a=-0.1 for n=100

framework in Section 5.5.2, but now with one more term, gives us

$$\begin{aligned} L(u) &= L(u_0)+L'(u_0)(u-u_0)+\frac{1}{2}L''(u_0)(u-u_0)^2+\frac{1}{6}L^{(3)}(u)(u-u_0)^3 \\ &\quad +\frac{1}{24}L^{(4)}(u^*)(u-u_0)^4 \end{aligned} \tag{6.2}$$

where $u^*$is between u_0 and u.

Taking the expectation on both sides of equation (6.2), gives us

$$\begin{aligned} E[L(u)] &= L(u_0)+L'(u_0)E(u-u_0)+\frac{1}{2}L''(u_0)E[(u-u_0)^2]+\frac{1}{6}L^{(3)}(u)E[(u-u_0)^3] \\ &\quad +\frac{1}{24}L^{(4)}(u^*)E[(u^*-u_0)^4] \end{aligned}$$

Recall that in the development of the Taylor based method we considered u_0 as fixed and was placed outside the expectation. We used

$$\begin{aligned} u_0 &= \hat{\mu}_{n,M_i}-\hat{\mu}_{n,wide} \\ u &= \hat{\mu}_{n,M_i}-\mu_{true} \end{aligned}$$

	$\widehat{EL}^{Taylor}_{n=1000,a=0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=0.1}$		$\widehat{EL}^{Taylor}_{n=1000,a=-0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=-0.1}$	
$M_{\{0\}}$	266.8230	18	1480.2040	18	461351.3417	18	1484.9103	18
$M_{\{01\}}$	4.6798	6	4.7742	6	4.7568	6	4.7593	6
$M_{\{02\}}$	260.9830	15	1449.2147	16	445895.2670	16	1447.2954	16
$M_{\{03\}}$	265.4117	17	1473.2340	17	456994.1585	17	1472.7860	17
$M_{\{04\}}$	56.5258	12	71.6674	12	137.9891	12	68.1847	12
$M_{\{05\}}$	262.1130	16	1448.4557	15	437699.6611	14	1433.1074	14
$M_{\{012\}}$	0.6440	2	0.6812	2	0.6378	2	0.6733	2
$M_{\{013\}}$	4.0991	5	4.1702	5	4.1783	5	4.1809	5
$M_{\{015\}}$	1.3562	4	1.3958	4	1.3329	4	1.3674	4
$M_{\{023\}}$	259.4185	13	1441.6026	14	441676.2191	15	1435.5024	15
$M_{\{024\}}$	50.0444	9	62.3635	8	111.6152	8	59.9403	8
$M_{\{034\}}$	55.6312	11	70.4296	11	135.1434	11	67.4214	11
$M_{\{035\}}$	260.4798	14	1440.4579	13	433240.7490	13	1420.6103	13
$M_{\{045\}}$	50.9189	10	63.6722	10	116.1080	10	61.5662	10
$M_{\{0123\}}$	0.0757	1	0.0926	1	0.0757	1	0.0924	1
$M_{\{0135\}}$	0.7298	3	0.7562	3	0.7251	3	0.7490	3
$M_{\{0234\}}$	49.1410	7	61.1448	7	109.1914	7	59.2186	7
$M_{\{0345\}}$	50.0042	8	62.4166	9	113.3003	9	60.6986	9

Table 6.8: Estimated LINEX loss for $F^{-1}(0.05)$, b=100, a=0.1 and a=-0.1 for n=1000

Ignoring the remainder, we see that for the second order Taylor approximation to underestimate, the third order Taylor term must be positive, and for second order Taylor approximation to overestimate, the third order term must be negative.

A major source for a substantive third order term is a large deviation between the error estimates u and u_0. Recall that u is estimated via the FIC framework, while u_0 is estimated directly. In our experiment it can be shown by example that the FIC framework is likely to be relative poor in estimating the error on the tail. We have picked a random focus from our sample from the n=1000 simulation. In Table 6.9 and Table 6.10, we have listed the focus, and the error estimated directly and via the FIC framework for $E(Y_{new} \mid x_j)$ and $F^{-1}_{Y_{new}}(0.05 \mid x_j)$. We see that while the two estimates of errors are perfectly aligned for $E(Y_{new} \mid x_j)$, there is a large deviation between the estimates of the two methods for $F^{-1}_{Y_{new}}(0.05 \mid x_j)$, especially with respect to the poorest models. In fact, we see that the FIC framework gives exactly the same error estimates for $F^{-1}_{Y_{new}}(0.05 \mid x_j)$. This is not surprising however, taken into accout Remark 6.1, telling us that the gradient ω are the same for the two foci. Consequently, this shows that we must be very careful when making strong inferences based on FIC for tail-focuses for poor models. In fact, we can observe that the various models produce very different estimated standard deviations, which is likely to affect the estimates of the differences $\hat{\mu}_{n,M_i} - \mu_{true}$. This can be seen in Table 6.1 as commented upon in Remark 6.4.

	$\widehat{\mu}_{n=1000}$	$\widehat{BIAS}_{FIC,n=1000}$	$\widehat{\mu}_{n=1000} - \widehat{\mu}_{n=1000,wide}$
$M_{\{0\}}$	10.443	-16.737	-16.737
$M_{\{01\}}$	26.858	-0.322	-0.322
$M_{\{02\}}$	9.306	-17.875	-17.875
$M_{\{03\}}$	11.866	-15.314	-15.314
$M_{\{04\}}$	30.517	3.336	3.336
$M_{\{05\}}$	9.306	-17.874	-17.874
$M_{\{012\}}$	24.778	-2.402	-2.402
$M_{\{013\}}$	29.313	2.132	2.132
$M_{\{015\}}$	25.477	-1.703	-1.703
$M_{\{023\}}$	10.743	-16.437	-16.437
$M_{\{024\}}$	28.348	1.168	1.168
$M_{\{034\}}$	32.821	5.640	5.640
$M_{\{035\}}$	10.816	-16.365	-16.365
$M_{\{045\}}$	29.066	1.886	1.886
$M_{\{0123\}}$	27.265	0.085	0.085
$M_{\{0135\}}$	28.038	0.858	0.858
$M_{\{0234\}}$	30.687	3.507	3.507
$M_{\{0345\}}$	31.485	4.305	4.305
$M_{\{012345\}}$	27.180	0.000	0.000

Table 6.9: Estimated focus E(Y) and error based on FIC and directly for a randomly picked covariate combination from the sample

We see that the extremeness of the second order Taylor development are even more exaggerated for the more extreme focus $F^{-1}_{Y_{new}}(0.01)$. When analyzing the corresponding focus on the other tail of the distribution, namely $F^{-1}_{Y_{new}}(0.95)$, the results are very similar that to $F^{-1}_{Y_{new}}(0.05)$, except that Taylor method now overestimate for positive a, and underestimate for negative a. The results for $F^{-1}_{Y_{new}}(0.01)$ and $F^{-1}_{Y_{new}}(0.95)$ are listed in Appendix B.

A caution to be noted is that it is easy to conclude from above that the direct method results are likely to be superior to the Taylor-based results, since the Taylor-based results are so extreme. This is likely to be true, but still a warning is in order. The direct method results based on the FIC framework cannot be better than the FIC-results themselves. Actually we see, when comparing Table 6.6 and Table 6.8, that the estimated expected losses for the focus $F^{-1}_{Y_{new}}(0.05)$ are suspiciously close to the $E(Y_{new})$ results, which is natural since the FIC-results are very close, cf. remark 6.1. However, our analysis in Remark 6.4 above reveals the imperfection of FIC in for the tail focus. The underlying FIC values are all based on the wide model estimate of σ, which is substantially lower than the σ-estimates when important covariates are missing. This is likely to understate the true MSE. At least the Taylor-based method somehow captures the impreciseness of FIC.

	$\widehat{\mu}_{n=1000}$	$\widehat{BIAS}_{FIC,n=1000}$	$\widehat{\mu}_{n=1000} - \widehat{\mu}_{n=1000,wide}$
$M_{\{0\}}$	-38.344	-16.737	-57.143
$M_{\{01\}}$	17.091	-0.322	-1.707
$M_{\{02\}}$	-39.419	-17.875	-58.218
$M_{\{03\}}$	-36.910	-15.314	-55.709
$M_{\{04\}}$	11.739	3.336	-7.060
$M_{\{05\}}$	-39.360	-17.874	-58.159
$M_{\{012\}}$	16.199	-2.402	-2.599
$M_{\{013\}}$	19.708	2.132	0.909
$M_{\{015\}}$	16.681	-1.703	-2.118
$M_{\{023\}}$	-37.971	-16.437	-56.770
$M_{\{024\}}$	10.238	1.168	-8.560
$M_{\{034\}}$	14.116	5.640	-4.682
$M_{\{035\}}$	-37.838	-16.365	-56.637
$M_{\{045\}}$	10.842	1.886	-7.956
$M_{\{0123\}}$	18.876	0.085	0.077
$M_{\{0135\}}$	19.440	0.858	0.641
$M_{\{0234\}}$	12.656	3.507	-6.143
$M_{\{0345\}}$	13.346	4.305	-5.453
$M_{\{012345\}}$	18.799	0.000	0.000

Table 6.10: Estimated focus $F^{-1}(0.05)$ and error based on FIC and directly for a randomly picked covariate combination from the sample

A rather surprising result at first sight, is that if we look at the direct estimates, the asymmetry in the loss function does not seem to have a large impact on the estimated expected loss, even though the error of the models are likely to be large. One should imagine the impact of large errors would be heavily affected by asymmetric loss functions. However, if we think about what we are really estimating, this is not so surprising. Recall that we average over all covariates in estimating the expected loss function. This means that we are averaging over covariates for which the errors of the focus are likely to go in different directions, and what we get is the average effect. If we instead estimated the focus for one particular covariate combination, asymmetry is likely to affect estimated expected loss. This will be illustrated next.

We used the the random covariate of the n=1000 simulation as used above, and estimated the expected loss for the focus $F^{-1}_{Y_{new}}(0.05 \mid x_j)$ associated with this particular covariate combination, using only the direct estimation method this time to avoid the additional problems when using the Taylor-based estimates. The estimated expected losses are given in Table 6.11. In Table 6.11, we see that the asymmetry of the loss function has a high impact on the estimated expected loss. We will get back to the consequences of this in the discussion below. This table also gives us an oppurtunity to produce an estimated expected loss

	$\widehat{EL}^{Direct}_{n=1000,a=0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=-0.1}$	
$M_{\{0\}}$	86.1296	16	265.9716	16
$M_{\{01\}}$	0.0850	1	0.0883	1
$M_{\{02\}}$	95.4913	18	318.9099	18
$M_{\{03\}}$	74.7840	13	209.7746	13
$M_{\{04\}}$	6.2928	9	5.0223	9
$M_{\{05\}}$	95.4861	17	318.8606	17
$M_{\{012\}}$	2.7002	8	3.1838	8
$M_{\{013\}}$	2.5711	7	2.2031	7
$M_{\{015\}}$	1.4041	5	1.5822	5
$M_{\{023\}}$	83.7191	15	253.6136	15
$M_{\{024\}}$	0.7588	4	0.6953	4
$M_{\{034\}}$	19.5573	12	13.3546	12
$M_{\{035\}}$	83.1323	14	250.5672	14
$M_{\{045\}}$	1.9445	6	1.7050	6
$M_{\{0123\}}$	0.1137	2	0.1118	2
$M_{\{0135\}}$	0.4935	3	0.4544	3
$M_{\{0234\}}$	7.0940	10	5.5676	10
$M_{\{0345\}}$	10.9185	11	8.1383	11

Table 6.11: Estimated LINEX loss for $F^{-1}(0.05)$ for a randomly picked covariate combination from the sample as focus, b=100, a=0.1 and a=-0.1 for n=1000

versus estimated focus plot for $F^{-1}_{Y_{new}}(0.05 \mid x_j)$. We see from Figure 6.2 the worst models produce very deviant foci, i.e. those models missing x_1 or its substitute x_4. In fact even the sign of foci are different and we see that the estimated expected loss is very high, especially for negative a.

Another observation from table 6.11 is that for this particular focus we get another winner model. While averaging resulted in the true model, $M_{\{0123\}}$, to be picked as the winner model, the simple model, $M_{\{01\}}$, is picked as the winner model. We see that the direction of the asymmetry does not affect model selection, as for the averaged results above. We might expect more variations in model selection when we look at single covariates, and an analysis of an additional randomly picked covariate combination confirms this. The results for such a new covariate combination are given in table 6.12.

In table 6.12 we see that the asymmetry does not affect the two winner models. However, the asymmetry cause a disagreement from the third best model. Hence, the asymmetry affect model selection. We also see that for this particular covariates the less precise alternative to x_2, x_5, provides a slightly lower estimated expected loss than the more precise alternative. This shows that even at n=1000 we are not guaranteed that the "correct" covariates are selected.

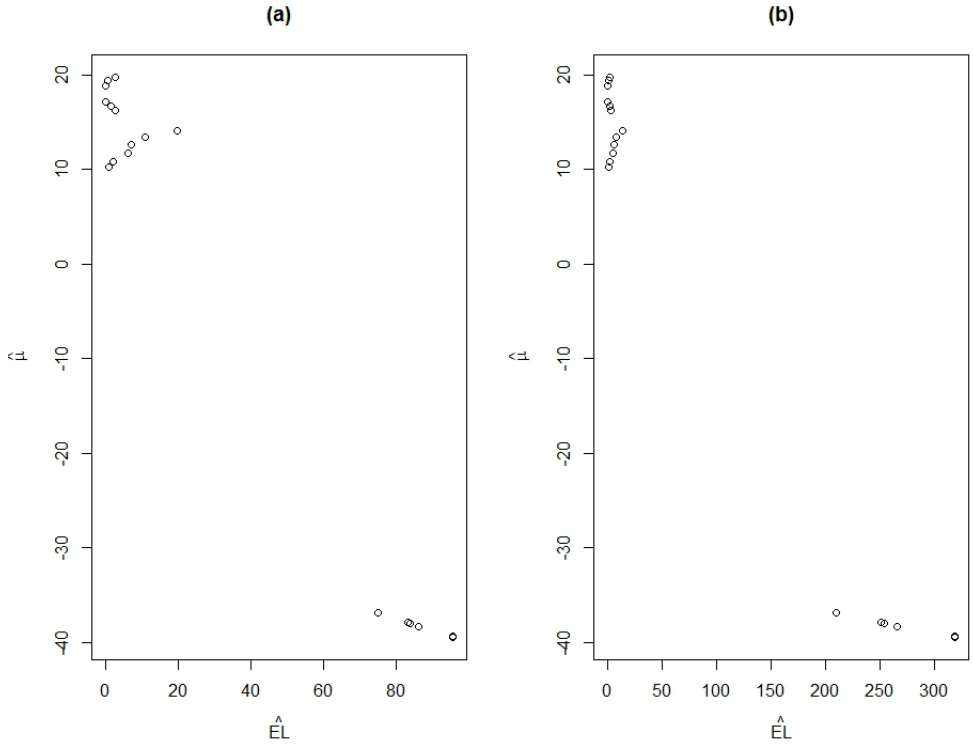

Figure 6.2: In part (a), we have plotted $\widehat{EL}^{Direct}_{n=1000,a=0.1}$ against estimated focus $\hat{\mu}_{n=1000}$ for focus $F^{-1}_{Y_{new}}(0.05 \mid x_j)$ for a randomly picked covariate. and in part (b) we have plotted the same, but for negative a

	$\widehat{EL}^{Direct}_{n=1000,a=0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=-0.1}$	
$M_{\{0\}}$	1555.1438	17	202.2527	17
$M_{\{01\}}$	1.4269	5	1.2725	3
$M_{\{02\}}$	1276.2196	14	186.7498	14
$M_{\{03\}}$	1727.8888	18	210.6879	18
$M_{\{04\}}$	34.3126	11	21.0763	11
$M_{\{05\}}$	1144.8244	13	178.4616	13
$M_{\{012\}}$	1.2929	4	1.4538	5
$M_{\{013\}}$	5.5193	6	4.4558	6
$M_{\{015\}}$	1.1954	3	1.3391	4
$M_{\{023\}}$	1422.6986	16	195.1745	16
$M_{\{024\}}$	8.4557	7	6.5233	7
$M_{\{034\}}$	51.7388	12	28.8774	12
$M_{\{035\}}$	1283.6860	15	187.1724	15
$M_{\{045\}}$	8.7237	8	6.7041	8
$M_{\{0123\}}$	0.0886	2	0.0902	2
$M_{\{0135\}}$	0.0870	1	0.0869	1
$M_{\{0234\}}$	16.8005	9	11.7725	9
$M_{\{0345\}}$	17.4372	10	12.1448	10

Table 6.12: Estimated LINEX loss for $F^{-1}(0.05)$ for another randomly picked covariate combination from the sample as focus, b=100, a=0.1 and a=-0.1 for n=1000

6.4.3 The information value of x_3

As for the analysis of the other foci above, the model $M_{\{012\}}$ is ranked as number one when n=100. The estimated expected loss increases by including x_3. Hence, x_3 should not be gathered whatever cost. When n=1000, the estimated expected loss decreases when including x_3. The information value of including x_3 seems to not be to a noticeable extent be influenced by the method used for estimation or asymmetry of the loss function. As for the focus $E(Y_{new})$ the information value of x_3 seems to be around 0.57 independent of the estimation method and the value of a. Hence, even if the estimated loss is slightly higher higher compared to the focus $E(Y_{new})$, this does not affect the information value of x_3. However, taking into account Remark 6.1 above, telling us that the two foci are invariant with respect to FIC values for normal linear models, except for a constant, this is not surprising at all.[74]

An interesting observation from Table 6.11, however, is that the information value of x_3, given that we know the particular covariate combination to use as a focus, has changed. In fact, for this particular focus, nothing is gained from gathering x_3. However, if we look at the other randomly picked covariate illustrated in Table 6.12, x_3 has a value of about 1.20 for a=0.1 and more than 1.3 for a=-0.1, which are

[74] See example 6.6 in Claeskens and Hjort (2008) p. 163.

larger than in the average case. A reasonable speculation is therefore that if we have gathered x_1 first, the information value of gathering x_3 might be altered. This illustrates the benefit of using sequential decision making in what covariates to gather. As indicated in Chapter 2 sequential decision-making can be optimized. For instance, it must be decided what covariate to gather first. In theory optimal sequential decision-making may be obtained by backward induction. This may be complex in practice, and other algorithms have been developed as described in Chapter 2. It will be beyond the scope of this study to elaborate further on this, but we note that the possibility of sequential covariate gathering may substantially improve the cost information-value trade-off when gathering covariates.

6.4.4 The value of more precise information

As for the other foci studied above, x_4 hardly seems to be a good substitute to x_1 unless c_1 is extremely costly to gather (relative to c_4). As for $E(Y_{new})$, however, gathering x_5 instead of x_2 seems to be a more practical alternative, especially when n=100. In the case of n=100, gathering x_5 instead of x_2 seems to only increase the estimated expected loss by about 0.20 when using $M_{\{024\}}$ instead of the winner model $M_{\{012\}}$. If we, when n=1000, substitute $M_{\{0123\}}$ by $M_{\{0135\}}$ the estimated expected loss increases by approximately 0.65, corresponding to the same number found for the focus $E(Y_{new})$. Hence, changing focus in this case doesn't seem to alter the value of more precise information.

6.4.5 The value of increased amount of data

We see that increasing n form 100 to 1000 has quite substantial effect on the estimated expected loss. This effect is not crucially dependent on the method used or the value of a. If we use the direct estimate for $a = 0.1$ as an example we see that for the winner model, $M_{\{012\}}$, the estimated expected loss is 1.05, while the estimated expected loss is 0.09 for the winner model, $M_{\{0123\}}$, when n=1000. If we compare to the analysis the focus $E(Y_{new})$, we seem that increasing n seems to pay-off more for $F_Y^{-1}(0.05)$. This is not surprising as more data is genrally likely to pay more off at the margin when estimating tail-related foci.

6.5 Chapter summary

Setting up a good simulation experiment is not an easy task.[75] The main objective of this simulation experiment was to illustrate and compare the different estimation methods discussed in this study for foci reasonable comparable. Our experiment was useful, both in showing that different methods, where applicable, produce fairly similar results, as they should do, but also revealing some dangers and pitfalls associated with the different methods.

[75] For practical guidelines and some of the challenges to set up a simulation experiment see Boos and Stefansky (2013) Chapter 9.

One particular lesson from this experiment was that it clearly provided some warnings about using the FIC framework in estimating the expected loss. We must pay careful attention to the underlying assumptions of the FIC framework. The framework seems to work well for foci associated with the center of the distribution. However, the method produced poor results for foci associated with the tail of the distribution of a new response, especially when working with very wrong models. This became especially apparent when the FIC-framework was used in combination with a second order Taylor development. Hovewer, the direct methods also has its problems related to the underlying assumptions of FIC.

Our analysis didn't produce any spectacular results showing how different foci or loss functions may substantially alter model selection and information-value of covariates, although marginal effects were discovered. This is not very surprising since we averaged over covariates, and this will even out certain features associated with certain covariate combinations. When returning to the analysis of single covariates more spectacular results were found. More spectacular results could also be obtained by other foci and different methods, but then the results would be less comparable. One "spectacular" result would, for instance, be that in the particular simulation setup above, and if our focus was the standard deviation, σ, FIC would not provide guidance in selecting model as no covariates are included in this parameter. FIC would be the same for all models. Hence, no covariate would have any information-value. The consequence of this would be that it would not be cost-efficient to gather costly covariates. The reason for this, as pointed out in Remark 6.4, is that asymptotically, σ is estimated equally well in all submodels. A natural extention of this study would be to perform other experiments that are able to push the theory more towards its limits. For instance, an interesting extention would be to test the methods on other categories of GLM regression.

One interesting topic briefly explored, but beyond the scope of this study to fully explore, would be to apply sequential decision making as described in Chapter 2 in gathering covariates. Let us say that we, for instance, find that x_1 is worth to gather. When x_1 is first gathered, this can provide valuable information with respect to the other covariates to gather. It is possible to imagine that for instance for some values of x_1, x_3 is worth to gather, while for others not. The estimation methods explored in this study are compatible with sequential covariate gathering. For instance, when x_1 is first gathered we could considered it as a protected variable in a FIC sense, and then perform a new analysis on what additional covariates to gather. Hence, the general theoretical estimation framework explored in this study with be easy applicable in settings where we gather the covariates sequentially.

7 Concluding remarks and more things to do

In this study we have explored several methods to estimate the expected loss associated with the use of regression models. The purpose was to find the information-value of covariates to be traded off against covariate costs. This can be used to assess which of costly covariates to gather in the estimation or prediction of a focus associated with a new combination of covariates.

We found that for the purpose for prediction of an observable variable, cross-validation is a generally applicable method compatible with most regression settings and loss functions. For particular settings such that the normal linear regression setting and quadratic loss, we can find proper bias correction terms to the plug-in empirical estimate.

For the estimation setting we developed one general method based on second order Taylor development of the loss function and the use of FIC. This method is generally applicable for smooth loss functions of estimation error. Our simulation experiment revealed, however, that this approach must be used with care. The approximation is not likely to be good for poor models, especially when operating with foci associated with the distribution tail. For some loss functions, such as the LINEX loss function, a direct approach is possible, and it is advised to explore such approaches when possible to avoid the problems associated with Taylor approximations. However, those methods are not perfect, either, as we have seen. Careful attention must be given to the underlying assumptions of the FIC-framework, and in partular the consequence of these assumptions when working with very wrong models. In this context it is also worth mentioning that FIC is based on first order Taylor developments itself. A further development of FIC, for instance, by second order Taylor developments could potentially improve its use in the Taylor development of a general loss function.

We have suggested several aspects to be explored further throughout this study. We will not repeat all those aspects here. Here we will only address a few major aspects that should be explored further at a summary level.

Firstly, in this study we have only been concerned with obtaining an unbiased estimate or approximately unbiased estimate for the expected loss associated with a model. We have not systematically discussed the quality of such an estimate, except briefly discussed how the data partitioning might affect the quality of a cross-validation estimate. Finding estimates is just the start of statistical inference. A natural extension to this study would be the estimation of the variance of the estimates. This could be used to compare methods and search for best possible method where many methods are available. Monte Carlo sampling can be used to estimate the variance for experimental models, while bootstrapping can be used to estimate the variance of estimates for a specific data set. Figure 6.1, where AIC was calculated for 100 different samples of the same underlying DGP, is an example on how Monte Carlo methods can be used to estimate variances for AIC. Variance reducing measures would be a natural supplement to the study of unbiased estimates.

Secondly, in this study we have not elaborated on sequential decision making in gathering covariates, and did not explore this systematically in the experiment. This could potentially be useful, because when a covariate is gathered, the estimation of loss for alternative models given that this covariate is gathered may be changed. For instance when experiencing that a covariate of a new case is of an extreme value, this might trigger different covariate selection choices compared to a covariate with a normal value. However, the estimation methods explored in in study applies equally well in a sequential covariate gathering framework. As mentoned in Chapter 2, Bayesian theory on optimal statistical decisions provides a theoretical framework for how such sequential covariate gathering can be optimally performed, and there

exists algorithms that can be performed to implement optimal or near optimal covariate gathering.

Furthermore, a topic we have not touched in this study at all, is the issue of model-averaging. Often, predictions and estimations may be improved by model-averaging as described in Claeskens and Hjort (2008) Chapter 7. Such model-averaging is compatible with the idea of this study. Assume that one covariate is gathered. Then it will not cost more to average over different models including this covariate. Hence, the information value of covariates may be altered by taking into account model-averaging.

Also, as we have seen in this study, FIC is well compatible with the LINEX loss function. The LINEX loss is mathematically desirable and flexible to fit a waste of real world loss function. As this was only a small part of the study, we didn't have the opportunity to go into more theoretical aspects and investigations regarding combining FIC and LINEX. We do however see that this is a promising area for further research. Since FIC gives us a limit distribution of the difference between an estimated focus and the true focus, i.e. the error, it may also be applied in combination with other loss functions relying on this error. One such possibility is the zero-one loss function, where the loss is dependent on whether the absolute error exceeds some ε.

Finally, in the application of the FIC-framework in this study, we have seen that careful attention must be given to the underlying assumptions of the FIC-framework. We generally think that it that other experiments, pushing the FIC-framework to its limits, would be useful in finding directions for the further development of the FIC-framework.

Bibliography

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Pages 267–281 in B. N. Petrov, and F. Csaki, (eds.) *Second International Symposium on Information Theory. Akademiai Kiado*, Budapest.

Anderson, D.R (2007). *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistical Surveys* 4 (2010):40-79.

Beneplanc, G. and Rochet, J-C. (2011). *Risk Management in Turbulent Times*. Oxford University Press.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition. Springer.

Boltzmann, L. (1877). Uber die Beziehung zwischen dem Hauptsatzederzwe: Ten mechanischen Warmetheorie und der Wahrscheinlichkeitsrechnung respective den Satzen uber das Warmegleichgewicht. *Wiener Berichte* 76, 373–435.

Burnham, K.P., and Anderson, D.R (2002). *Model selection and multimodel inference : a practical information-theoretic approach*. Springer.

Boos, D. D. and Stefansky, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. Springer.

Casella, G. and Berger, R.L. (2001), *Statistical Inference*. Second Edition. Duxbury Thomson Learning

Claeskens, G. and Hjort, N.L. (2003). The focused information criterion. *Journal of the American Statitical Association*, 98:900-916. With discussion and rejoinder by the authors.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill.

DeGroot, M. H. (1984). Changes in utility as information. *Theory and Decision* 17: 287–303.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapmann and Hall. London.

Friedman, J. H. (1997). On Bias, Variance, 0/1-Loss and the Curse-of-Dimensionality. *Data mining and Knowledge Discovery 1:55*. Kluwer Academic Publishers.

Gujarati, D. (2002). *Basic Econometrics.* 4 edition. McGraw-Hill/Irwin.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer.

Hjort, N. L. and Claeskens, G. (2003). Rejoinder to "The focussed information criterion" and "Frequentist model average estimators". Journal of the American Statistical Association, 98, 938-945.

Hjort, N. L. and Claeskens, G. (2008). Minimizing Average Risk in Regression Models. *Econometric Theory*, 24, 493-527.

Knight, K (2000). *Mathematical Statistics*. Chapman & Hall/CRC Texts in Statistical Science.

Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Second edition, Chapman and Hall, London.

Neumann J. von and Morgenstern O. (1944). *Theory of Games and Economic Behaviour*. Princeton, NJ: Princeton University Press.

Parmigiani, G and Inoue, L. (2009). *Decision Theory: Principles and Approaches*. Wiley.

Polansky, A. M (2011). Introduction to Statistical Limit Theory. CRC Press

Posner, R. A. (2010). *Economic Analysis of Law*. Eight edition. Aspen Casebooks.

Raiffa, H. And Schlaifer R. (1961). *Applied Statistical Decision Theory*. Harward University Press.

Savage, L.J. (1954). *The Foundations of statistics*, John wiley and Sons Inc.

Schwarz, G. (1978). *Estimating the dimension of a model*. Annals of Statistics 6, 461–464.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal 27, 379–423 and 623–656.

Varian, H. R. (1992). *Microeconomic Analysis*. Third edition. W. W Norton & Company.

Wassermann, L. (2003). *All of statitsics: A concise course course in statistical inference*. Springer

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. Suri-Kagaku (Mathematic Sciences) 153, 12–18. (In Japanese.Wasserman, L., *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.

Varian, H.(1974). A Bayesian approach to real estate assessment, in: S. E. Feinberg and A. Zellner, eds., Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage. (North Holland: Amsterdam) 195-208.

Witten I. H,, Eibe, F., Hall, M.A. (2011). *Data mining : practical machine learning tools and techniques*. 3rd ed. Morgan Kaufmann.

Young, G. A. and Smith, R. L. (2005), *Essentials of Statistical Inference*. Cambridge University Press.

Zellner, A. (1986): "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," Journal of the American Statistical Association, 81, 446-451.

A Selected R issues

We will not provide the full R code for the simulation experiment in this study. This will involve a lot of code without much value for the reader without proper guidance. All R code may, however, be provided upon request to the author.

In this appendix we will provide some fragments of the R code that may be useful for the reader for carrying out the analysis covered by this study. We will provide generic code for finding numerical derivatives of functions. We will show how to numerically estimate the Fisher information matrix. We will show how one relatively simply automated can organize models with various combinations of covariates to be run in a loop. This, inter alia, is useful when calculating information values and other measurements for many models at once. We will show an easy algorithm for estimating leave-one-out cross-validation expected loss function. Next, we present some some useful code to execute the FIC-analysis. Finally, we will provide the seed and R code used to generate the data for the experiment. This allows the reader to replicate the experiment.

We will assume that the readers of this appendix are familiar with R. R is intuitive in its design, though, which makes it possible for readers not familiar with R (but some basic insight in programming) to read most of the code as pseudo-code.

A.1 Numerical derivatives

R has has a built in function "deriv" to find derivatives of a function. For practical purposes it is often convenient to manually find the gradient of a function numerically, i.e the partial derivatives. The approximation done in this study is to use the formula that for small ε we have

$$f'(x) \approx \frac{f(x+\varepsilon)+f(x-\varepsilon)}{2\varepsilon}$$

The R code for implementing this approximation goes as follows

```
getgrad = function(objective,para) {
eps=10^(-6)
I.n=diag(length(para))
grad.mu = vector(length=length(para))
for (i in 1:length(para)) {
grad.mu[i]=(do.call("objective",list(para+eps*I.n[,i]))-
do.call("objective",list(para-eps*I.n[,i])))/(2*eps) }
return(grad.mu)      }
```

The function "getgrad" takes a generic function and finds the gradient of the "objective" at parameter values given by "para". In the code we have set $\varepsilon = 10^{-6}$

A.2 Estimating the Fisher information matrix

Estimating the Fisher information matrix, i.e. finding $\hat{J}_n$, is crucial to statistical inference. Finding $\hat{J}_n$ is crucial for FIC analysis, where we basically everything flows from estimating $\hat{J}_{n,wide}$. We will return to FIC analysis below. In the normal regression setting it is not difficult manually estimate $\hat{J}_n$, as this is given by

$$\hat{J}_n = \frac{1}{\hat{\sigma}_n} \begin{bmatrix} 2 & 0 \\ 0 & X^t X/n \end{bmatrix}$$

However, in more complicated setting, numerical calculations might be necessary. An easy approach is to use "nlm" or other optimization functions in R. Typically $\hat{J}_n$ can be obtained by the following R code

```
res=nlm(minusloglik, startpoint, hessian=T)
mle=res$estimate
j=res$hessian/n
```

"minusloglik" is here the negative of

$$\ell_n(\theta) = \log \mathscr{L}_n(\theta) = \sum_{i=1}^{n} \log f(y_i;\theta)$$

We need to use the negative of the log likelihood function, since nlm minimize a function. "startpoint" is some vector of numbers given to nlm as a starting point for the iterative minimization procedure, and should ideally be cleverly chosen. The nlm function is typically used to find the MLE numerically as illustrated in the code. A nice property of nlm is that it returns the Hessian, H, at the MLE, which can be used to estimate $\hat{J}_n$. By WLLN and the properties of MLE, we have that

$$\frac{H}{n} = \hat{J}_n \xrightarrow{p} J$$

A.3 Automated organization of models

When we want to assess and/or analyze several models it soon becomes useful to store the models in a list that allows us to run through all models and extract relevant measurements in a loop. Even if we run through relatively few models in the experiment of this study, this approach is time saving. In the following R code, we show how we stored the 19 analyzed models with their associated covariates in a list.

```
ones=rep(1,length(y))
protected=ones
free=cbind(x1,x2,x3,x4,x5)
subsets=rbind(
```

```
c(0,0,0,0,0),
c(1,0,0,0,0),
c(0,1,0,0,0),
c(0,0,1,0,0),
c(0,0,0,1,0),
c(0,0,0,0,1),
c(1,1,0,0,0),
c(1,0,1,0,0),
c(1,0,0,0,1),
c(0,1,1,0,0),
c(0,1,0,1,0),
c(0,0,1,1,0),
c(0,0,1,0,1),
c(0,0,0,1,1),
c(1,1,1,0,0),
c(1,0,1,0,1),
c(0,1,1,1,0),
c(0,0,1,1,1),
c(1,1,1,1,1))
models=list()
for (i in 1:nrow(subsets))
models[[i]]=cbind(protected,free[,subsets[i,]==1])
modelnames=c(0,01,02,03,04,05,012,013,015,023,024,
034,035,045,0123,0135,0234,0345,012345)
```

First we generate a vector of "ones" being the covariates associated with the protected constant term. x1 to x5 are the free covariates that may or may not be included in the different models. "subsets" is a matrix where the number rows corresponds to the number of models and the columns represent the coding of each model. Each row code the different models. 1 corresponds to the inclusion of a covariate, while 0 means that a covariate is not included. For instance the second row, c(1,0,0,0,0), codes the models where only x1 is included in addition to the protected constant term.

We have coded all the models manually since we will only analyze a selection of models. An alternative would be to use the function "expand.grid" in R, which would automatically generate a matrix all combinations of 1 and 0 for the length we want. For instance, all 32 possible models could be generated by the following R code

```
subsets=expand.grid(0:1,0:1,0:1,0:1,0:1)
```

To store all the models we create a list named "models". Then we run a loop where the covariates in all the models corresponding to the codes in subsets are stored. For instance, models[[2]] would contain

a matrix with a column of the protected "ones" , and the second column consisting x1. This can be used to run through a loop analyzing and storing relevant measurements for all the models. For instance in our case the R code

```
modres=list()
for (i in 1:length(models))
modres[[i]]=normalregression(y,models[[i]])
```

run a loop that creates a list "modres" storing the relevant measurements each model by the pre-determined function "normalregression". "normalregression" will typically generate measurements such as MLE, AIC, $\hat{J}_n$ for regression model with response "y" and the covariates in models[[i]]. For instance, we can extract the AIC for model 2 by

```
modres[[2]]$aic
```

"modelnames" is simply a vector of the name of each model in the order of subsets, that make each model easily recognizable when presenting relevant information for each model

A.4 Doing leave-one-out cross-validation

Doing cross-validation is essential to estimate the expected loss functions associated with the alternative models. Here we will show a simple algorithm to estimate the expected loss for all the alternative models in this study by leave-one-out cross validation. The R code goes as follows:

```
lvcvest=vector(length=length(models))
for (j in 1:length(models)) {
lvcvest[j]=0
for (i in 1:ndata) {
ynew=y[-i]
xnew=as.matrix(models[[j]][-i,])
newbetas = normalregression(ynew,xnew)$mle[-1]
lvcvest[j]=lvcvest[j]+QL(y[i]-models[[j]][i,]%*%newbetas)      }
lvcvest[j]=lvcvest[j]/ndata
}
```

The first loop, indexed by j, runs through all the models. lvcvest is a vector of the same length as the number of models to store the estimate for each model. Initially this is set to zero. The inner loop indexed by i is running through all the observations for each model. We sequentially take out one observation, and the remaining observations are used to estimate β by MLE. The loss for the observation in question is obtained by calling an external loss function, taking the difference between the observed value and estimated value as an argument. Here, QL, means that we use the quadratic loss, coded as follows:

```
QL = function (x) x^2
```

QL could easily be replaced by another function, for instance LINEX.

The sum of all losses is divided by the number of observations in the end to obtain the average loss associated with the model in question.

A.5 Executing the FIC analysis

Although the theory and calculation behind FIC might appear complex at first sight, the actual execution of FIC analysis is not too complex.[76]

Firstly, we must decide upon the candidate models to be considered. Relevant model information, from the wide model to the most narrow model can be stored in a list as described in A.3. Secondly, although the focus itself is not necessary to do FIC analysis, we will normally be interested in the value of the focus, or the foci to be averaged over, for all models. In some cases it might be instructive to see how the focus varies with the models in a graph.

We can then start with the FIC analysis. Most of the information needed can be extracted from estimates associated with the wide model. From $\hat{J}_{n,wide}$ and its inverse we can extract the relevant sub-matrices as presented in the R code code below.

```
pn=2 ## Number of protected parameters
qn=5 ## Number of "free" parameters
wideres=normalregression(y,models[[length(models)]])
Jwidehat =wideres$j
Jinvhat = solve(Jwidehat)
Q=Jinvhat[(pn+1):(pn+qn),(pn+1):(pn+qn)]
J00=Jwidehat[(1:pn),(1:pn)]
J10=Jwidehat[(pn+1):(pn+qn),1:pn]
J01=Jwidehat[1:pn,(pn+1):(pn+qn)]
J11=Jwidehat[(pn+1):(pn+qn),(pn+1):(pn+qn)]
```

Having estimates of the relevant matrices, we can go on. We need the MLE for the wide model, which we have stored from before as described in section A.3. Furthermore, we need an estimate of the gradient vector for the focus μ under the wide model,

$$\widehat{\nabla}\mu_{n,wide} = \begin{bmatrix} \frac{\partial\mu}{\partial\theta}\Big|_{\theta=\hat{\theta}_{n,wide}} \\ \frac{\partial\mu}{\partial\gamma}\Big|_{\gamma=\hat{\gamma}_{n,wide}} \end{bmatrix}$$

The gradient estimate might be found numerically as described in A.1. Having this, we now have the

[76] See Claeskens and Hjort (2008) p. 153 for a similar guidance. The present guidance is however more specific with regards to the R code while Claeskens and Hjort (2008) provides more mathematical details.

relevant input to obtain the estimates $\hat{\tau}^2_{0,n}$, $\hat{\omega}^t_n$, and $\hat{\delta}_{n,wide}$, which are common for all models in finding FIC. This is done in the following R code:

```
widemle=wideres$mle
gradmuwide=grad.mu(widemle,x0.model[[length(models)]])
omegahat=J10%*%solve(J00)%*%gradmuwide[1:pn]-gradmuwide[(pn+1):(pn+qn)]
tau0sqhat=t(gradmuwide[1:pn])%*%solve(J00)%*%gradmuwide[1:pn]
deltahat=sqrt(length(y))*(widemle[(pn+1):(pn+qn)]-seq(0,0,length.out=qn))
```

We can now turn to the model-specific measurements, i.e the measurements that must be done for each model to obtain FIC. In FIC analysis the π-matrices are essential to extract the model specific information from the Fisher information matrix of the wide model, $J_{wide.}$. In the following R code we generate a list of the π-matrices associated with each model.

```
pival=list()
pival[[1]]=0
for (j in 2:nrow(subsets)){
where = (1:qn)[subsets[j, ] == 1]
pival[[j]]=diag(qn)[where,]
if (is.vector(pival[[j]]))
pival[[j]]=t(as.matrix(pival[[j]])) }
```

First we establish that the π-matrix of the model with none of the free parameters/covariates are simply zero. To generate the remaining π-matrices we utilize the subsets as explained and described in A.3. Having the necessary π-matrix for each model it is quite easy to find the model specific measurements for each model used to calculate FIC. For the technical details we refer to Section 3.3.2. In the following R-code, we have created a loop that runs through all models to find FIC and associated measurements

```
for (i in 1:length(models)) {
if (i==1) GS=0*Iq else      {
Iq=diag(qn)
QS=solve(pival[[i]]%*%solve(Q)%*%t(pival[[i]]))
QSnull=t(pival[[i]])%*%QS%*%pival[[i]]
GS=QSnull%*%solve(Q)        }

var=tau0sqhat+t(omegahat)%*%GS%*%Q%*%t(GS)%*%omegahat
bsq=t(omegahat)%*%(Iq-GS)
%*%(deltahat %*% t(deltahat)-Q)%*%t(Iq-GS)%*%omegahat
FIC=var+max(bsq,0)
BIAS=t(omegahat)%*%(deltahat-GS%*%deltahat)/sqrt(ndata)
MSE=FIC/ndata
keep[i,c(2:6)]=c(var,BIAS,bsq,FIC,MSE)
```

For the purposes of our study we have created a matrix "keep", that stored estimates of the variance, bias, bias squared, FIC and MSE for each model. The first column of keep is reserved to store the μ estimate associated with each model.

The above guidance was related to finding FIC for a single focus. For the purposes of this study, we are are concerned with averaging the FIC over all the covariates in the sample, or averaging over the foci using AFIC. When averaging over all the covariates in the sample we can simply create a loop where we run through all the covariates, in each step setting the a new covariate in the sample as the focus. The FIC for each focus and other relevant can be stored in a matrix to be averaged and assessed. A disadvantage with this method that for a large sample and many models this is likely to take some time to process, as there will be a chain of more than one loop. For the purposes of this study, however, the processing of finding FIC for all models for all covariates never took longer time than the time needed to make a cup of coffee. When AFIC is used for averaging it is necessary to compute the measurements for AFIC. We will not show this code here, but may, as mentioned be provided upon request to the author.

A.6 The simulated data

The code below describes the simulated data used for the simulation experiment. The only variable not defined is ndata which in the experiment is the number of simulated data, n=100 and n=1000.

```
beta0=10
beta1=10
beta2=1
beta3=1
sigma=5
sigmax3=1
noise=1

set.seed (10)
x1=runif(ndata,-5,5)
x2=runif(ndata,-5,5)
x3=rnorm(ndata,0,sigmax3)
x4=x1+rnorm(ndata,0,noise)
x5=x2+rnorm(ndata,0,noise)
y=beta0+beta1*x1+beta2*x2+beta3*x3+rnorm(ndata,0,5)
```

The random covariate was picked by the following code

```
set.seed(3)
f=sample(1:ndata,1)
```

f index the covariate picked randomly from the number of observations in the sample. For the second randomly covariate, we used the same code, but set the seed to 5.

B Experimental values for focus $F^{-1}_{Y_{new}}(0.01)$ and $F^{-1}_{Y_{new}}(0.95)$

B.1 Experimental results for $F^{-1}_{Y_{new}}(0.01)$

The results for estimating $F_Y^{-1}(0.01)$ for n=100 and n=1000 are given in Table B.1 and Table B.2, respectively.

	$\widehat{EL}^{Taylor}_{n=100,a=0.1}$		$\widehat{EL}^{Direct}_{n=100,a=0.1}$		$\widehat{EL}^{Taylor}_{n=100,a=-0.1}$		$\widehat{EL}^{Direct}_{n=100,a=-0.1}$	
$M_{\{0\}}$	26.0482	9	884.0283	16	1578788.1676	18	1253.2518	17
$M_{\{01\}}$	2.9498	5	3.5182	5	3.0903	5	3.5495	5
$M_{\{02\}}$	24.6106	7	863.6604	13	1539306.8926	14	1252.6987	16
$M_{\{03\}}$	28.6100	12	900.8545	18	1540928.9503	15	1236.6043	13
$M_{\{04\}}$	30.0363	16	47.0289	12	145.8083	12	53.8251	12
$M_{\{05\}}$	25.3506	8	878.9720	14	1575635.1553	17	1255.1247	18
$M_{\{012\}}$	0.9740	1	1.2554	1	0.9993	1	1.2500	1
$M_{\{013\}}$	3.1251	6	3.6062	6	3.2969	6	3.6171	6
$M_{\{015\}}$	1.1395	3	1.4624	3	1.1751	3	1.4503	3
$M_{\{023\}}$	27.0965	10	880.1618	15	1514427.3485	13	1245.3073	15
$M_{\{024\}}$	30.9199	18	46.7194	10	126.2585	9	48.0856	8
$M_{\{034\}}$	28.8326	13	44.6672	8	136.2864	11	53.2269	11
$M_{\{035\}}$	27.9448	11	896.1645	17	1541931.3405	16	1241.3989	14
$M_{\{045\}}$	30.8724	17	46.9116	11	130.1528	10	49.3514	10
$M_{\{0123\}}$	1.1226	2	1.3569	2	1.1463	2	1.3509	2
$M_{\{0135\}}$	1.3107	4	1.5687	4	1.3499	4	1.5589	4
$M_{\{0234\}}$	29.6091	14	44.3503	7	118.1184	7	47.2749	7
$M_{\{0345\}}$	29.6864	15	44.6825	9	121.8069	8	48.4945	9

Table B.1: Estimated LINEX loss for $F^{-1}(0.01)$, b=100, a=0.1 and a=-0.1 for n=100

B.2 Experimental results for $F^{-1}_{Y_{new}}(0.95)$

The results for estimating $F_Y^{-1}(0.01)$ for n=100 and n=1000 are given in Table B.3 and Table B.4, respectively.

	$\widehat{EL}^{Taylor}_{n=1000,a=0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=0.1}$		$\widehat{EL}^{Taylor}_{n=1000,a=-0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=-0.1}$	
$M_{\{0\}}$	20.0238	12	1480.4816	18	5580764.3473	18	1485.1887	18
$M_{\{01\}}$	4.6238	6	4.7926	6	4.8792	6	4.7778	6
$M_{\{02\}}$	18.3958	8	1449.4869	16	5380374.6385	16	1447.5672	16
$M_{\{03\}}$	19.6109	11	1473.5104	17	5525543.6130	17	1473.0623	17
$M_{\{04\}}$	40.0621	18	71.6975	12	346.5704	12	68.2143	12
$M_{\{05\}}$	19.0336	10	1448.7278	15	5269312.1211	14	1433.3768	14
$M_{\{012\}}$	0.6608	2	0.6988	2	0.6564	2	0.6910	2
$M_{\{013\}}$	4.0637	5	4.1885	5	4.2688	5	4.1992	5
$M_{\{015\}}$	1.3706	4	1.4136	4	1.3545	4	1.3852	4
$M_{\{023\}}$	17.9321	7	1441.8735	14	5326981.2963	15	1435.7722	15
$M_{\{024\}}$	36.2036	15	62.3920	8	261.2332	8	59.9684	8
$M_{\{034\}}$	39.4834	17	70.4595	11	336.6014	11	67.4508	11
$M_{\{035\}}$	18.5520	9	1440.7285	13	5212946.9148	13	1420.8775	13
$M_{\{045\}}$	36.6702	16	63.7009	10	274.8348	10	61.5946	10
$M_{\{0123\}}$	0.0932	1	0.1102	1	0.0933	1	0.1100	1
$M_{\{0135\}}$	0.7462	3	0.7739	3	0.7441	3	0.7667	3
$M_{\{0234\}}$	35.6027	13	61.1731	7	253.1657	7	59.2466	7
$M_{\{0345\}}$	36.0792	14	62.4451	9	265.5931	9	60.7268	9

Table B.2: Estimated LINEX loss for $F^{-1}(0.01)$, b=100, a=0.1 and a=-0.1 for n=1000

	$\widehat{EL}^{Taylor}_{n=100,a=0.1}$		$\widehat{EL}^{Direct}_{n=100,a=0.1}$		$\widehat{EL}^{Taylor}_{n=100,a=-0.1}$		$\widehat{EL}^{Direct}_{n=100,a=-0.1}$	
$M_{\{0\}}$	117289.1449	17	882.0309	16	331.4166	15	1250.5049	17
$M_{\{01\}}$	2.7941	5	3.3080	5	2.8195	5	3.3393	5
$M_{\{02\}}$	112591.5669	13	861.7043	13	333.5351	18	1249.9529	16
$M_{\{03\}}$	118072.4879	18	898.8229	18	327.2383	13	1233.8912	13
$M_{\{04\}}$	69.6514	12	46.7304	12	45.0531	12	53.5128	12
$M_{\{05\}}$	116384.1423	15	876.9848	14	331.7919	16	1252.3740	18
$M_{\{012\}}$	0.7842	1	1.0498	1	0.7807	1	1.0445	1
$M_{\{013\}}$	3.0162	6	3.3959	6	2.9791	6	3.4068	6
$M_{\{015\}}$	0.9551	3	1.2565	3	0.9494	3	1.2444	3
$M_{\{023\}}$	113446.6963	14	878.1722	15	332.2092	17	1242.5765	15
$M_{\{024\}}$	67.0883	10	46.4216	10	40.2538	8	47.7850	8
$M_{\{034\}}$	65.3332	9	44.3735	8	44.9516	11	52.9159	11
$M_{\{035\}}$	117259.3038	16	894.1424	17	328.5218	14	1238.6760	14
$M_{\{045\}}$	67.8214	11	46.6134	11	41.3382	10	49.0483	10
$M_{\{0123\}}$	0.9360	2	1.1512	2	0.9249	2	1.1451	2
$M_{\{0135\}}$	1.1336	4	1.3625	4	1.1176	4	1.3528	4
$M_{\{0234\}}$	63.0493	7	44.0573	7	39.8885	7	46.9760	7
$M_{\{0345\}}$	63.9464	8	44.3888	9	40.9325	9	48.1931	9

Table B.3: Estimated LINEX loss for $F^{-1}(0.95)$, b=100, a=0.1 and a=-0.1 for n=100

	$\widehat{EL}^{Taylor}_{n=1000,a=0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=0.1}$		$\widehat{EL}^{Taylor}_{n=1000,a=-0.1}$		$\widehat{EL}^{Direct}_{n=1000,a=-0.1}$	
$M_{\{0\}}$	459980.3310	18	1480.2040	18	267.9157	18	1484.9103	18
$M_{\{01\}}$	4.7708	6	4.7742	6	4.6656	6	4.7593	6
$M_{\{02\}}$	446452.6514	16	1449.2147	16	260.5310	16	1447.2954	16
$M_{\{03\}}$	457124.6165	17	1473.2340	17	265.3068	17	1472.7860	17
$M_{\{04\}}$	142.9081	12	71.6674	12	53.3499	12	68.1847	12
$M_{\{05\}}$	442085.1713	14	1448.4557	15	258.5203	15	1433.1074	14
$M_{\{012\}}$	0.6451	2	0.6812	2	0.6368	2	0.6733	2
$M_{\{013\}}$	4.1676	5	4.1702	5	4.1099	5	4.1809	5
$M_{\{015\}}$	1.3602	4	1.3958	4	1.3288	4	1.3674	4
$M_{\{023\}}$	443437.1604	15	1441.6026	14	257.9896	14	1435.5024	15
$M_{\{024\}}$	114.8152	8	62.3635	8	47.8053	8	59.9403	8
$M_{\{034\}}$	139.3608	11	70.4296	11	52.8833	11	67.4214	11
$M_{\{035\}}$	438901.5260	13	1440.4579	13	255.8295	13	1420.6103	13
$M_{\{045\}}$	118.9110	10	63.6722	10	48.9776	10	61.5662	10
$M_{\{0123\}}$	0.0758	1	0.0926	1	0.0756	1	0.0924	1
$M_{\{0135\}}$	0.7321	3	0.7562	3	0.7228	3	0.7490	3
$M_{\{0234\}}$	111.7169	7	61.1448	7	47.3581	7	59.2186	7
$M_{\{0345\}}$	115.5685	9	62.4166	9	48.4175	9	60.6986	9

Table B.4: Estimated LINEX loss for $F^{-1}(0.95)$, b=100, a=0.1 and a=-0.1 for n=1000

C Selected theorems and proofs

In this appendix we will present selected theorems and sketches of proofs fundamental to some of the most crucial results in this study. As the transition from IID random variables to a regression is explained in the text, we will only present the proofs for the IID context here. Furthermore, we will, as in most mainstream literature, restrict our proof to the one parameter case, an discuss the extension to the multiparameter case.

C.1 Asymptotic properties of the MLE

C.1.1 Assumptions and regularity conditions

As just mentioned we assume that Y_i, i=1,2..,n are IID random variables generated by the probability distribution $f(y;\theta)$. The likelihood function is given by

$$\ell_n(\theta) = \log \mathscr{L}_n(\theta) = \sum_{i=1}^{n} \ell(y_i;\theta)$$

where $\mathscr{L}_n(\theta) = \prod_{i=1}^{n} f(y_i;\theta)$ and $\ell(y_i;\theta) = \log f(y_i;\theta)$. The MLE, $\hat{\theta}_n$, found by maximizing $\ell_n(\theta)$ with respect to θ, is assumed to be the solution of $\ell_n'(\hat{\theta}_n) = 0$.

Many nice results can be derived if $f(y;\theta)$ is a probability density satisfying the regularity conditions commonly referred to in statistics. The regularity conditions in statistics have many different costumes in different texts. An intuitively appealing version of regularity assumptions in the one-parameter case are those presented in Knight (2000) p. 256 as:

(A1)	The parameter space Θ is an open subset of the real line.
(A2)	The set $A = \{x : f(x;\theta) > 0\}$ does not depend on θ.
(A3)	$f(x;\theta)$is three times continuously differentiable with respect to θ for all x in A.
(A4)	$E[\ell'(Y_i,\theta)] = 0$ for all θ and $VAR[\ell'(Y_i,\theta)] = K(\theta)$ where $0 < K(\theta) < \infty$ for all θ.
(A5)	$E[\ell''(Y_i,\theta)] = -J(\theta)$ for all θ where $0 < J(\theta) < \infty$ for all θ.
(A6)	For each θ and $\delta > 0$, $\lvert\ell'''(Y_i,\theta)\rvert \le M(y)$ for $\lvert\theta - t\rvert \le \delta$ where $E[M(Y)] < \infty$.

Note $E[\ell'(Y_i,\theta)] = 0$ will normally follow from (A2). This can be seen by noting that

$$\int_A f(y;\theta)dy = 1$$

Deriving both side gives us, as long as we can move derivative inside the integral,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int_A f(y;\theta)dy \\ &= \int_A \frac{\partial}{\partial\theta} f(y;\theta)dy \\ &= \int_A \ell'(Y_i,\theta)\, f(y;\theta)dy \\ &= E[\ell'(Y_i,\theta)] \end{aligned}$$

Also note that if we can integrate $\int_A f(y;\theta)dy$ twice inside the integral sign, we have $K(\theta) = J(\theta)$, because

$$\begin{aligned} 0 &= \int_A \frac{\partial}{\partial\theta}[\ell'(Y_i;\theta)\, f(y;\theta)]dy \\ &= \int_A \ell''(Y_i;\theta)\, f(y;\theta)dy + \int_A (\ell'(Y_i;\theta))^2 f(y;\theta)dy \\ &= -J(\theta) + K(\theta) \end{aligned}$$

Finally, note that the assumption above are too strict for our case as we will normally assume that the data are generated by an unknown DGP, $g(y)$, and that the density $f(y;\theta)$ is an approximation model for inference. Model selection, crucial to this study, is to choose among alternative densities according to their merits, or as in this study, to find information value of covariates. In line with the mainstream literature first assume that the data are generated by a known $f(y;\theta)$ and discuss the extension to the more general case where the data are generated by an unknown DGP.

C.1.2 Consistency of the MLE

Theorem. *Under assumptions (A1-A6) above,* $\hat{\theta}_n \xrightarrow{p} \theta$.

Proof. (High level sketch)

Rigorously proving consistence of the MLE is more subtle than one might initially think. We will only present a high level sketch of proof. This high level sketch of proof can be made short and elegant without using K-L distance, as the one presented in Casella and Berger (2001) p. 470. However, we find it instructive to relate the proof to the K-L distance as presented in Knight (2000) p. 260 and Wasserman (2003) p. 126. The reason is that using K-L distance makes it easier to understand the extension to the least false property of the MLE as presented in the text.

Let θ_0 now denote the true value of θ. Then note that maximizing $\ell_n(\theta)$ with respect to θ is equivalent to maximizing

$$\Phi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} log\frac{f(y_i;\theta)}{f(y_i;\theta_0)}$$

since $\Phi_n(\theta) = \frac{1}{n}(\ell_n(\theta) - \ell_n(\theta_0))$ and $\ell_n(\theta_0)$ is a constant. Now, observe that

$$\begin{aligned} E(log\frac{f(Y_i;\theta)}{f(Y_i;\theta_0)}) &= \int_A log\frac{f(y_i;\theta)}{f(y_i;\theta_0)} f(y_i;\theta_0) \\ &= -\int_A log\frac{f(y_i;\theta_0)}{f(y_i;\theta)} f(y_i;\theta_0) \\ &= -D(\theta_0,\theta) \end{aligned}$$

By WLLN, $\Phi_n(\theta) \xrightarrow{p} -D(\theta_0,\theta)$. Since $-D(\theta_0,\theta) \leq 0$ (can be proved with Jensen's inequality) with $-D(\theta_0,\theta_0) = 0$, we would expect that $\Phi_n(\theta)$ tends to be maximized at θ_0, hence, that $\hat{\theta}_n$ tends to θ_0. To prove this formally, we need to prove that $\Phi_n(\theta) \xrightarrow{p} -D(\theta_0,\theta)$ implies that $\hat{\theta}_n \xrightarrow{p} \theta_0$ uniformly over θ. We will not go into the details of this here, but refer to the neat proof in Wasserman (2003) p. 135. □

The extension to the multiparameter case is not very complex, and the proof is for practical purposes the same, but assuming θ to be a vector. The principle of the proof is also to a large extent valid for the situation where θ_0 is not the true value of θ, but the least false parameter in the sense that it minimizes $-D(\theta_0,\theta)$, with the modification that $f(y;\theta)$ does not necessarily correspond to the true DGP, $g(y)$.

C.1.3 Asymptotic normality of the MLE

Theorem. *Under assumptions (A1-A6) above,* we have that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \frac{K(\theta)}{J(\theta)^2})$$

Proof. (Sketch based on the proof in Knight (2000) p. 263)

The essence of the proof is that we have that the MLE is the solution to

$$\ell_n'(\hat{\theta}_n) = \sum_{i=1}^{n} \ell'(y_i;\hat{\theta}_n) = 0$$

A first order Taylor development of this term gives us

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \ell'(y_i;\hat{\theta}_n) \\ &= \sum_{i=1}^{n} \ell'(y_i;\theta) + (\hat{\theta}_n - \theta)\sum_{i=1}^{n} \ell''(y_i;\theta) \\ &\quad + \frac{1}{2}(\hat{\theta}_n - \theta)^2 \sum_{i=1}^{n} \ell'''(y_i;\theta_n^*) \end{aligned}$$

where θ_n^* is between θ and $\hat{\theta}_n$. Multiplying both sides by $\sqrt{n}$ gives us

$$\begin{aligned} 0 &= \sqrt{n}\sum_{i=1}^{n}\ell'(y_i;\theta)+\sqrt{n}(\hat{\theta}_n-\theta)\sum_{i=1}^{n}\ell''(y_i;\theta) \\ &\quad +\sqrt{n}\frac{1}{2}(\hat{\theta}_n-\theta)^2\sum_{i=1}^{n}\ell'''(y_i;\theta_n^*) \end{aligned}$$

Solving this gives us

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n-\theta) &= \frac{-\sqrt{n}\sum_{i=1}^{n}\ell'(y_i;\theta)}{\sum_{i=1}^{n}\ell''(y_i;\theta)+\sqrt{n}\frac{1}{2}(\hat{\theta}_n-\theta)^2\sum_{i=1}^{n}\ell'''(y_i;\theta_n^*)} \\ &= \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(y_i;\theta)}{-\frac{1}{n}\sum_{i=1}^{n}\ell''(y_i;\theta)-\frac{1}{\sqrt{n}}\frac{1}{2}(\hat{\theta}_n-\theta)^2\sum_{i=1}^{n}\ell'''(y_i;\theta_n^*)} \end{aligned}$$

There are several ways to proceed from here. One way is to recognize by the CLT that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell'(y_i;\theta)=\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i;\theta)\xrightarrow{d}N(0,K(\theta))$$

and by the WLLN that $-\frac{1}{n}\sum_{i=1}^{n}\ell''(y_i;\theta)\xrightarrow{p}J(\theta)$, and finally, from the consistency of MLE $\frac{1}{\sqrt{n}}\frac{1}{2}(\hat{\theta}_n-\theta)^2\sum_{i=1}^{n}\ell'''(y_i;\theta_n^*)\xrightarrow{p}0$. Then the Theorem follows by applying Slutsky's theorem. □

First note that for must practical purposes assuming $f(y;\theta)$ being the true density of the DGP, we have that $K(\theta)=J(\theta)$. Consequently,

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{d}N(0,J(\theta)^{-1})$$

This is, in fact, the most common representation of the asymptotic normality of MLE in textbooks.

Furthermore, note that the proof of the multiparameter case follows the same lines. The parallel to $\frac{K(\theta)}{J(\theta)^2}$ becomes $J(\theta)^{-1}K(\theta)J(\theta)^{-1}$, which reduces to $J(\theta)^{-1}$ when $K(\theta)=J(\theta)$.

Finally, note that the extension to the case where the DGP, $g(y)$, will follow the same principles as here. But in this case, we will not have $K(\theta)=J(\theta)$.

C.2 Invariance of the MLE

Theorem. *Let $\hat{\theta}_n$ be the MLE. Then $\hat{\mu}_n = \mu(\hat{\theta}_n)$ is the MLE of $\mu(\theta)$ for functions μ*

Proof. (as presented in Wasserman (2003) p. 128)

Let $h = \mu^{-1}$. Then $\hat{\theta}_n = h(\hat{\mu}_n)$. For any μ, $\mathscr{L}_n(\mu) = \prod_{i=1}^{n} f(y_i; h(\mu)) = \prod_{i=1}^{n} f(y_i; \theta) = \mathscr{L}_n(\theta)$ where $\theta = h(\mu)$. Hence, for any μ, we have $\mathscr{L}_n(\mu) = \mathscr{L}_n(\theta) \leq \mathscr{L}_n(\hat{\theta}_n) = \mathscr{L}_n(\hat{\mu}_n)$ □

Note that a more rigorous proof based on the notation $\hat{\theta}_n = \underset{\theta}{sup}\mathscr{L}_n(\theta)$ can be found in Casella and Berger (2001) p. 320.

C.3 The delta method

Theorem. *Let Y_n be a sequence of random variables such that $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. For a given function μ and a specific value of θ suppose that $\mu'(\theta)$ exists and is not zero. Then*

$$\sqrt{n}(\mu(Y_n) - \mu(\theta)) \xrightarrow{d} N(0, \mu'(\theta)^2\sigma^2)$$

Proof. (Sketch based on Casella and Berger (2001) p. 243 and Knight (2000) p. 132)

A first order Taylor expansion of $\mu(Y_n)$ around $Y_n = \theta$ gives

$$\mu(Y_n) = \mu(\theta) + \mu'(\theta)(Y_n - \theta) + R_n$$

Which can also be written as

$$\mu(Y_n) - \mu(\theta) = \mu'(\theta)(Y_n - \theta) + R_n$$

Consequently, by applying Slutsky's theorem, we have that

$$\sqrt{n}(\mu(Y_n) - \mu(\theta)) = \sqrt{n}(Y_n - \theta)(\mu'(\theta) + \frac{R_n}{(Y_n - \theta)}) \xrightarrow{d} N(0, \mu'(\theta)^2\sigma^2)$$

Since $\frac{R_n}{(Y_n - \theta)}$ is $o_p(1)$. See Knight (2000) p. 132 for more details. □

The proof the multiparameter case is similar, but it must be taken into account that σ^2 is replaced by a covariance matrix C, which means that $\mu'(\theta)^2\sigma^2$ must be replaced by $\nabla\mu^t C \nabla\mu$

C.4 Elements of the FIC framework (FICology)

It will be beyond the scope of this study to presents proofs or even sketches of proof for the entire FIC-framework. However, to give the reader an idea of the proof, we will present sketches of proof of a simplified framework assuming only two alternative models: a wide and a narrow. Hence, we will follow the framework in Claeskens and Hjort (2008) Chapter 5.

C.4.1 The framework

In the FIC framework it is assumed that Y has the density

$$f_n(y) = f(y; \theta_0, \gamma_0 + \delta/\sqrt{n})$$

θ_0 is a vector of those parameters that are always included (protected parameters) and is assumed to be of dimension p. $\gamma = \gamma_0 + \delta/\sqrt{n}$ represent the free parameters, which is of dimension q. The subset of models $\{M_i\}_{i=1,\ldots,m}$ compromise the various models between the full (wide) model and the narrowest model. In our simplified framework we will assume two alternative models, the wide models which includes all the free parameters, and the narrow where none of the free parameters are included. We let $\hat{\theta}_{narr}$ be the MLE for θ_0 under the narrow model, while $\hat{\theta}_{wide}$ is the MLE estimator/estimate for θ_0 under the wide model. Furthermore $\hat{\gamma}_{wide}$ is the MLE under the wide model.

Let the score function for the wide model be

$$\begin{aligned} \ell'(y; \theta_0, \gamma_0) &= \begin{bmatrix} \frac{\partial log\, f(y;\theta_0,\gamma_0+\delta/\sqrt{n})}{\partial \theta} \\ \frac{\partial log\, f(y;\theta_0,\gamma_0+\delta/\sqrt{n})}{\partial \gamma} \end{bmatrix} \\ &= \begin{bmatrix} \ell'(y;\theta_0) \\ \frac{\partial log\, f(y;\theta_0,\gamma_0+\delta/\sqrt{n})}{\partial \gamma} \end{bmatrix} \end{aligned}$$

Furthermore, let J be the Fisher information matrix for the wide model, which can be split into the protected and free parameters. Hence,

$$J = \begin{bmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{bmatrix} \text{ and } J^{-1} = \begin{bmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{bmatrix}$$

Finally, let $\mu_{true} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$ be the true value of the focus, $\hat{\mu}_{narr}$ the estimated focus under under the narrow model, and $\hat{\mu}_{wide}$, the estimated focus under the wide model using MLE. Hence,

$$\begin{aligned} \hat{\mu}_{narr} &= \mu(\hat{\theta}_{narr}, \gamma_0) \\ \hat{\mu}_{wide} &= \mu(\hat{\theta}_{wide}, \hat{\gamma}_{wide}) \end{aligned}$$

C.4.2 Limit distribution of parameters

Theorem. *In the limit we have that*

$$\begin{aligned} \sqrt{n}\begin{bmatrix} \hat{\theta}_{wide} - \theta_0 \\ \hat{\gamma}_{wide} - \gamma_0 \end{bmatrix} &\xrightarrow{d} N(\begin{bmatrix} 0 \\ \delta \end{bmatrix}, J^{-1}) \\ \sqrt{n}(\hat{\theta}_{narr} - \theta_0) &\xrightarrow{d} N(J_{00}^{-1} J_{01}\delta, J_{00}^{-1}) \end{aligned}$$

Proof. (Sketch based on Claeskens and Hjort (2008) p. 122) For the first part we have by the asymptotic normality of MLE

$$\sqrt{n}\begin{bmatrix} \hat{\theta}_{wide} - \theta_0 \\ \hat{\gamma}_{wide} - (\gamma_0 + \delta/\sqrt{n}) \end{bmatrix} \xrightarrow{d} N(0, J^{-1})$$

which can be rewritten as

$$\sqrt{n}\begin{bmatrix} \hat{\theta}_{wide} - \theta_0 \\ \hat{\gamma}_{wide} - \gamma_0 \end{bmatrix} - \begin{bmatrix} 0 \\ \delta \end{bmatrix} \xrightarrow{d} N(0, J^{-1})$$

which can be rewritten as the first part of the theorem.

When it comes to the second part of the theorem, we basically use that

$$\ell_n'(\hat{\theta}_{narr}) = \sum_{i=1}^{n} \ell'(y_i; \hat{\theta}_{narr}) = 0$$

and by using the same principles we used to prove the asymptotic normality of the MLE, we get that

$$\sqrt{n}(\hat{\theta}_{narr} - \theta_0) \doteq_p J_{00}^{-1} \sqrt{n} \frac{\sum_{i=1}^{n} \ell'(Y_i; \theta_0)}{n}$$

where $A_n \doteq_p B_n$ means that $A_n - B_n$ tends to zero in probability. Note that

$$\begin{aligned} f(y; \theta_0, \gamma_0 + \delta/\sqrt{n}) &= f(y; \theta_0, \gamma_0) + \left.\frac{\partial f}{\partial \gamma}\right|_{\gamma=\gamma_0} \frac{\delta}{\sqrt{n}} + R_n \\ &= f(y; \theta_0, \gamma_0)(1 + \left.\frac{\partial log\, f}{\partial \gamma}\right|_{\gamma=\gamma_0} \frac{\delta}{\sqrt{n}}) + R_n \end{aligned}$$

Using this, gives us

$$\begin{aligned} E(\ell'(Y_i; \theta_0) &= \int \ell'(Y_i; \theta_0) f(y; \theta_0, \gamma_0 + \delta/\sqrt{n}) dy \\ &= \int \ell'(Y_i; \theta_0)(f(y; \theta_0, \gamma_0)(1 + \left.\frac{\partial log\, f}{\partial \gamma}\right|_{\gamma=\gamma_0} \frac{\delta}{\sqrt{n}}) + R_n) dy \\ &\doteq_p \int \ell'(Y_i; \theta_0)(f(y; \theta_0, \gamma_0)(1 + \left.\frac{\partial log\, f}{\partial \gamma}\right|_{\gamma=\gamma_0} \frac{\delta}{\sqrt{n}}) dy \\ &= J_{01} \frac{\delta}{\sqrt{n}} \end{aligned}$$

since the R_n part is $o_P(\frac{1}{\sqrt{n}})$. The result then follows. □

C.4.3 Limit distribution of a focus

Theorem. *In the limit, we have that*

$$\begin{aligned}\sqrt{n}(\hat{\mu}_{narr} - \mu_{true}) &\xrightarrow{d} N(\omega^t \delta, \tau_0^2) \\ \sqrt{n}(\hat{\mu}_{wide} - \mu_{true}) &\xrightarrow{d} N(0, \tau_0^2 + \omega^t Q \omega)\end{aligned}$$

where

$$\begin{aligned}Q &= J^{11} = (J_{11} - J_{10} J_{00}^{-1} J_{01}) \\ \tau_0^2 &= \left(\frac{\partial \mu}{\partial \theta}\right)^t J_{00}^{-1} \frac{\partial \mu}{\partial \theta} \\ \omega &= J_{10} J_{00}^{-1} \frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}\end{aligned}$$

and the derivatives are taken at (θ_0, γ_0)

Proof. (Sketch). We will take the wide part first. By the delta method we have that

$$\sqrt{n}(\hat{\mu}_{wide} - \mu_{true}) \xrightarrow{d} N(0, \tau^2)$$

where

$$\begin{aligned}\tau^2 &= \begin{bmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{bmatrix}^t J^{-1} \begin{bmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{bmatrix}^t \begin{bmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{bmatrix} \begin{bmatrix} \frac{\partial \mu}{\partial \theta} \\ \frac{\partial \mu}{\partial \gamma} \end{bmatrix}\end{aligned}$$

By matrix manipulations we will find that

$$\begin{aligned}J^{01} &= -J_{00}^{-1} J_{01} Q \\ J^{00} &= J_{00}^{-1} + J_{00}^{-1} J_{01} J_{10} J_{00}^{-1} Q\end{aligned}$$

Inserting this in the expression of τ^2, we find that

$$\tau^2 = \tau_0^2 + \omega^t Q \omega$$

For the narrow part, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_{narr} - \mu_{true}) &= \sqrt{n}(\mu(\hat{\theta}_{narr}, \gamma_0) - \mu(\theta_0, \gamma_0 + \delta/\sqrt{n}) \\
&= \sqrt{n}(\mu(\hat{\theta}_{narr}, \gamma_0) - \mu(\theta_0, \gamma_0 + \delta/\sqrt{n}) \\
&= \sqrt{n}(\mu(\hat{\theta}_{narr}, \gamma_0) - \mu(\theta_0, \gamma_0)) - \sqrt{n}(\mu(\theta_0, \gamma_0 + \delta/\sqrt{n}) - \mu(\theta_0, \gamma_0)) \\
&\doteq_p \frac{\partial \mu}{\partial \theta}^t \sqrt{n}(\hat{\theta}_{narr} - \theta_0) - \sqrt{n}\frac{\partial \mu}{\partial \gamma}\frac{\delta}{\sqrt{n}} \\
&\xrightarrow{d} N(\omega^t \delta, \tau_0^2)
\end{aligned}$$

which proves the statement □

We have now presented sketches of proof for a simplified FIC framework assuming only a wide and narrow model. The FIC framework also includes all models "in between". The principles for deriving the results from this framework are the same, but with some complications as described in the main text.

D List of abbreviations

AIC	An Information Criterion/Akaikes Information Criterion
AFIC	Averegaed Focussed Information Criterion
AFICM	Averaged Focused Information Criterion Modified
CLT	Central Limit Theorem
DGP	Data Generating Process
EL	Expected Loss
FIC	Focused Information Criterion
GLM	Generalized Linear Model
IID/iid	Independent Identically Distributed
J	Fisher information matrix
K-L/KL	Kullback-Leibler
LINEX	Linear Exponential
MLE	Maximum Likelihood Estimator
MSE	Mean Squared Error
VaR	Value at Risk
WLLN	Weak Law of Large Numbers